Measuring coselectional constraint in learner corpora: A graph-based approach

Dissertation

zur Erlangung des akademischen Grades

Doktorin/Doktor der Philosophie (Dr. phil.)

eingereicht an der Sprach- und literaturwissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

M.A. Anna Valer'evna Shadrova

Prof. Dr. -Ing. Dr. Sabine Kunst
Präsidentin
der Humboldt-Universität zu Berlin

Prof. Dr. Ulrike Vedder
Dekanin
der Sprach- und
literaturwissenschaftlichen Fakultät

Gutachterinnen und Gutachter:

1. Prof. Dr. Anke Lüdeling (Humboldt-Universität zu Berlin)
2. Prof. Dr. Amir Zeldes (Georgetown University)

Datum der Einreichung: 07.01.2020

Datum der Disputation: 10.07.2020

To Theo, Sasha, Nura, Saif, and Yousef:

May curiosity be your blessing,
and awe and wonder your companion –
Whatever you find along the way:
May it open your hearts
as much as your minds.

Zusammenfassung

Die Dissertation behandelt die strukturelle Entwicklung von Koselektionsbeschränkungen von Verben und Argumentkopflexemen bei Lerner*innen des Deutschen. Der Begriff Koselektionsbeschränkung bezieht sich hier auf die viel beschriebene Neigung von Muttersprachler*innen, sich bei der gemeinsamen Auswahl von Wörtern und Wörtern mit syntaktischen Strukturen auf eine vergleichsweise geringe Menge von konventionell oder durch andere Prozess eingegrenzte Koselektionen zu beschränken (Firth, 1957; Erman and Warren, 2000; Yorio, 1989). Für diese Eigenschaft prägte Sinclair (1991) den Begriff des sog. *idiomatischen Prinzips* ('idiom principle'), welches innerhalb der gebrauchsbasierten Linguistik weitläufig als Eigenschaft natürlicher Sprache anerkannt ist (Granger, 2005; Nesselhauf, 2005; Ellis, 2008; Herbst, 2014a; Wray, 2002; Pawley and Syder, 1983; Pawley et al., 2007; Wulff, 2008; Goldberg, 2006, 1995; Michaelis, 2012). Gleichsam anerkannt ist, dass der Erwerb solcher Koselektionsbeschränkungen selbst für weit fortgeschrittene Lerner*innen einer Fremdsprache große Schwierigkeiten darstellt und nur selten zu einem L1-ähnlichen Gebrauch führt (Paquot, 2019, 2018; Wray, 2002, 2013; Granger and Meunier, 2008; Erman and Warren, 2000; Yorio, 1989; Wulff, 2008).

Diese Arbeit hat zum Ziel, diese Entwicklung im L2-Erwerb für Verben und Argumentkopflexeme empirisch nachzuvollziehen. Dafür werden die Lexeme von Subjekten, Akkusativ-, Dativ-, Genitiv-, Präpositionalobjekten, Infinitivergänzungen und Objektsätzen sowie Prädikativen behandelt. Die Studie basiert auf dem kleinen bis mittelgroßen Korpus Kobalt (Zinsmeister et al., 2012), das aus 151 thematisch einheitlichen argumentativen Essays von chinesischen und belarusischen[1] und chinesischen Lerner*innen des Deutschen sowie 20 von deutschen Muttersprachler*innen verfassten Vergleichstexten besteht.

Dabei trifft die Untersuchung auf zwei zentrale Herausforderungen: Zum einen wird die theoretische Modellierung von Koselektionsbeschränktheit, so wie sie aktuell in der gebrauchsbasierten Linguistik diskutiert wird, der Komplexität des Phänomens bislang nicht gerecht. Koselektion ist in mehrfacher Hinsicht ein Schnittstellenphänomen zwischen Lexikon und Syntax sowie zwischen individuellen Spracheigenschaften und emergenten Gruppeneigenschaften, *parole* und *langue* nach Saussure (1916/1983), und hat darüber hinaus Schnittflächen mit weitreichenden Aspekten von Semantik, Semiotik, Phonotaktik und Phonetik bzw. auditorischem Gedächtnis. Demgegenüber existiert in der gebrauchsbasierten Beschreibung bislang hauptsächlich die Annahme eines sog. *phraseologischen Kontinuums*, auf dem Koselektionsphänomene verschiedener Art als 'mehr oder weniger festgelegte Sprache' angeordnet werden. Dabei werden sehr unterschiedliche Phänomene von Chunks (vollständig fixierte, morphosyntaktisch unanalysierte und kommunikative Einheiten) bis zu abstrakten Syntaxkonstruktionen auf einer einzigen Dimension angeordnet, was wichtige Konzepte vermischt, deren Auseinandersetzung für ein tieferes Verständnis von funktionalen und strukturellen Aspekten von Koselektion entscheidend wäre.

[1] Die Bezeichnung *Weißrussland* entstand im Zusammenhang mit historischen Verschiebungen unter Einfluss des russischen Reiches, während die Eigenbezeichnung *Belarus* auf die Kiewer Rus zurückgeht. Diese Arbeit folgt der Eigenbezeichnung und verwendet daher die Begriffe *Belarus* und *belarusisch* (*Belarusian* auf Englisch).

Zum anderen ist die quantitative Methodik und Methodologie für die Messung und Analyse von Koselektion unterentwickelt und epistemologisch problematisch. Die gegenwärtige Forschung stellt im Wesentlichen zwei Ansätze für die quantitative Analyse von Koselektionsphänomenen bereit: Einerseits könnte die Betrachtung von Koselektionsbeschränkungen als Negativ zur Messung von lexikalischer und morphosyntaktischer Produktivität verstanden werden (Baayen, 2001; Zeldes, 2012, 2013b), was allerdings problematisch ist, da Koselektionsbeschränktheit wahrscheinlich nicht als komplementär zu Produktivität aufzufassen ist – beispielsweise sind Muttersprachler*innen sowohl produktiver als auch stärker koselektionsbeschränkt als Lerner*innen. Andererseits Modelle aus der inferentiellen Statistik frequentistischer Art, wobei Koselektionsbeschränkungen in lexikalischen Assoziationsmaßen als Abweichungen von der erwarteten bedingten Wahrscheinlichkeit des gemeinsamen Auftretens zweier Wörter (Konstruktionen, etc.) aufgefasst werden. Diese Modellierung ist problematisch, weil sie a) gegen Zufallsverteilungen misst, was gegenüber dem enormen kombinatorischen Potential von Wörtern keine nützliche Baseline ist; b) auf Faktorkombinationen, also Exemplaren, aufbaut, während das linguistische Modell von einer strukturellen Eigenschaft ausgeht, somit nicht von einer simplen Additivität von Einzelphänomenen; c) epistemologische Fragen aufwirft, indem sie (große) Korpora als Stichprobe aus einer Sprachpopulation annimmt, deren ontologischer Status allerdings selbst zweifelhaft ist; und d) weil sie den Bedarf an exakten Modellen und quantitativen Messungen aus kleinen, aber tief annotierten Datenmengen nicht deckt, wie er in der variationistischen Linguistik insbesondere bei der Betrachtung von nicht-kanonischen Varietäten stets besteht.

Diese theoretischen und methodischen Aspekte werden in der Arbeit auf Grundlage eines quasi-longitudinalen Forschungsdesign betrachtet, wobei Ergebnisse aus einem validierten c-Test (onDaF, Eckes (2017)) als Ordungsvariable für den annähernden Sprachstand der Teilnehmer*innen zugrunde gelegt werden. Aus der angenommenen inneren Dynamik der *Interlanguage* nach Selinker (1972) wird abgeleitet, dass der Erwerb von Koselektionsbeschränkungen bei Lerner*innen einer Reihe dialektisch interagierender Prozesse unterliegt, nämlich einer Diversifizierung, einer Randomisierung und einer Spezialisierung, woraus sich die Hypothese eines U-Kurven-förmigen Erwerbsverlaufs ergibt; und dass Lerner*innen insgesamt über den Erwerbsverlauf koselektionsbeschränkter werden, aber selbst auf hohen Erwerbsständen kein muttersprachenähnliches Niveau erreichen.

Zunächst wird dies innerhalb einer statistischen Analyse operationalisiert, die Ergebnisse sind jedoch aufgrund der kombinatorischen Eigenschaften von Lexemen und möglicherweise aufgrund der geringen Datenmenge unschlüssig, obwohl eine Grundähnlichkeit bei der lexikalischen Auswahl und im Thema besteht. Anschließend wird eine Operationalisierung in einem graphbasierten Modell vollzogen. Hierzu wird für verschiedene Subkorpora die Graphmetrik *Louvain-Modularität* (Blondel et al., 2008) berechnet, die als *Community-Detection*-Algorithmus die innere Strukturiertheit eines Graphs misst. Diese graphbasierte Analyse liefert unerwartet klare Ergebnisse, die den meisten Hypothesen entsprechen, und erlaubt darüber hinaus feine Differenzierungen von hohem linguistischem Wert: Es zeigen sich unterschiedliche Effekte für Objekt- und Subjektslots bei Lerner*innen gegenüber Muttersprachler*innen, textstrukturelle Effekte, und es ergeben sich einige Hinweise darauf, dass selbst bei kleinen Datenmengen in einem linguistisch eher simplistischen Graphmodell Unterschiede zwischen individuellen und Gruppeneffekten im Sinne von *langue* und *parole* beobachtbar sind. Das birgt erste Evidenz dafür, dass Graphmetriken als Operationalisierung für komplex verwobene linguistische Prozess selbst in kleinen Korpora wie Kobalt nützlich sind, während zugleich wichtige epistemologische

Problematiken umgangen werden können.

Damit leistet die Arbeit einen Beitrag zur Systematisierung von Methodenentwicklung und Methodologie in der Korpuslinguistik. Die eingeführte Graphmetrik wird umfassend gegen typische konfundierende Faktoren wie Korpusgröße oder Textlänge validiert, wofür Samplingverfahren eingeführt werden, die bislang in der Korpuslinguistik kaum Anwendung finden, nämlich ein Out-of-Sampling, ein Sliding-Window-Sampling und eine Textlängennormalisierung, die nicht auf der Tokenzahl, sondern auf der Zahl der gewünschten Strukturen (realisierte Verbargumentstrukturen) basiert. Eine tokenbasierte Normalisierung bleibt der strukturbasierten dabei unterlegen.

Abschließend werden die Ergebnisse aus typologischer, kultureller, individuell-kognitiver und lexikosyntaktischer Perspektive diskutiert, weiterführende Ideen für die Anwendung graphbasierter Methoden und Metriken in kernlinguistischen Forschungsfragen vorgestellt, und einige Anmerkungen zur Replikation und weiteren, insb. externen Validierung gemacht. Zuletzt werden Vorschläge für die Entwicklung eines funktionalen Modells von Koselektionsbeschränktheit skizziert.

Abstract

This thesis is about the development of coselectional constraint or nativelike coselection of verbs and argument lexemes in learners of German. Coselectional constraint here refers to the much described tendency of native speakers to limit their choices of words with other words or syntactic structures to a relatively small set of conventionalized or otherwise constrained coselections (Firth, 1957; Erman and Warren, 2000; Yorio, 1989). This has been coined the *idiom principle* by Sinclair (1991) and is widely accepted in usage-based linguistics as a property of natural language (Granger, 2005; Nesselhauf, 2005; Ellis, 2008; Herbst, 2014a; Wray, 2002; Pawley and Syder, 1983; Pawley et al., 2007; Wulff, 2008; Goldberg, 2006, 1995; Michaelis, 2012). It is also widely accepted that learners have a hard time acquiring such coselectional constraints even at highly advanced stages of acquisition (Paquot, 2019, 2018; Wray, 2002, 2013; Granger and Meunier, 2008; Erman and Warren, 2000; Yorio, 1989; Wulff, 2008).

The thesis seeks to verify this empirically for the coselection of verbs and their argument lexemes, viz. subject, accusative, dative, genitive, prepositional, infinitival, clause object, and predicate head lexemes. The study is based on the small to mid-sized, but tightly controlled and deeply annotated German learner corpus Kobalt (Zinsmeister et al., 2012), which contains 151 learner essays written in response to the same prompt by Belarusian[2] and Chinese learners of German, and 20 native speaker control texts.

In doing so, the analysis meets two major challenges: (1) The theoretical modeling of coselectional constraint as it exists today does not yet serve justice to the complexity of the phenomenon, which is situated at the very interface of lexicon and syntax, as well as community and individual language or *langue* and *parole* (Saussure, 1916/1983), and has a number of intersecting points with semantics, semiotics, phontactics, phonetics and auditory memory. In usage-based linguistics, however, coselectional constraint is usually subsumed under 'more or less fixed language' in a hypothesized monodimensional continuum ranging from chunks (entirely fixed, communicative bits of language that go morphosyntactically unanalyzed) to abstract syntax. This model conflates a number of concepts that are relevant to the deeper understanding of the structural and functional properties of coselectional constraint; and (2), the quantitative methodology for measuring coselection is underdeveloped and epistemologically problematic. At present, there are basically only two ways to analyze coselectional constraint quantitatively: Either through the analysis of productivity as a negative of coselectional constraint (Baayen, 2001; Zeldes, 2012, 2013b). This is problematic because coselectional constraint is likely not ideally modeled as complementary to productivity, since native speakers are both more productive and more coselectionally constrained; Or through frequentist models of inferential statistics, i.e. models of lexical association where constraint is defined as a higher than expected conditional probability of an item to occur with another one. These are problematic for a structural assessment of coselection, because they (a) test against randomness, which is a poor baseline against the massive combinatorial power of coselections; (b) rely on

[2]The terms *White Russia* or *Byelorussia* were coined in the context of historical shifts under involvement of the Russian Empire, while the self-designation Belarus goes back to the Kievan Rus. This work follows the endonym in using the terms *Belarus* and *Belarusian*.

factor combinations, i.e. exemplars, where the linguistic model presumes a structural phenomenon; (c) raise epistemological concerns in suggesting that a (large) corpus can validly be viewed as a sample from a language population of which the ontological status is in fact unclear; and (d), they do not satisfy the need of variationist and non-canonical linguistics for exact models and quantitative measurement in small, but deeply annotated data.

These theoretical and methodological aspects are addressed in a cross-sectional or 'quasi-longitudinal' study design, where results of a validated cloze-test (onDaF, Eckes (2017)) are used as an approximation of proficiency. It is hypothesized from the presumed inner workings of interlanguage (Selinker, 1972) that learners in acquiring coselectional constraints underly a number of dialectically interrelated and dynamic processes, namely diversification, randomization, and specialization, which is expressed in a u-shaped development; and that learners overall gain in coselectional constraint, but do not reach native-speaker levels even at very advanced stages of acquisition.

A first operationalization is attempted in a statistical analysis, but yields inconclusive results. This is in part due to the high combinatorial power of even a small corpus like Kobalt despite high lexical and thematic similarity between texts, and likely in part due to the relatively small size of the dataset. It is then translated to a graph-based model and an analysis based on a community detection algorithm which measures the inner structuredness of a graph, *Louvain modularity* (Blondel et al., 2008). This graph-based analysis in contrast yields unexpectedly clear results in line with most hypotheses, and allows for fine-grained differentiations of high linguistic value: Effects are different for subjects and objects in learners vs. native speakers; text-structural effects can be observed; and results tentatively suggest that some insight into the complex dynamic process of the emergence of grouped vs. individual effects can be gained even from a rather simplistic graph model. This provides a first evidence to the idea that graph metrics may serve as an operationalization for complexly interwoven linguistic processes even in small corpora like Kobalt, while also avoiding some of the epistemological caveats of text-based statistics.

The thesis also aspires to be a contribution to the systematization of methodological development in corpus linguistics. For this, the graph-based metric introduced is internally validated against typical confounding factors such as corpus size, text length, and group vs. individual effects by applying several sampling techniques to the data that are not yet common in the field, viz. an out-of-sampling, a sliding-window-sampling, and a text length normalization based on the structure relevant to the analysis (verb-argument structures), which proves superior to a token-based text length normalization.

Finally, results are discussed in light of typological, cultural, cognitive, and lexicosyntactic effects; some suggestions are made to the improvement of the linguistic model and the question of external validation; the more general issue of data size in corpus linguistics is discussed; some ideas for further application and development of graph-based models and metrics are outlined; and a tentative proposal for a functional description of coselectional constraint is presented.

Contents

Abstract (German) **1**

Abstract (English) **4**

Table of Contents **6**

List of Figures **9**

List of Tables **12**

Acknowledgements **13**

1. On fixedness, conventionality, and coselection **15**

2. Linguistic background **23**
2.1. Coselection in corpus studies 23
2.1.1. Collocations 24
2.1.2. Collostructional distribution 29
2.1.3. Idiosyncrasy 31
2.1.4. Summary 33
2.2. Coselection in language learning 34
2.2.1. Learner corpora, collocational competence and phraseological complexity 35
2.2.2. Distributional sensitivity 38
2.2.3. Coselection and interlanguage 43
2.2.4. Summary 49
2.3. A theory of coselectional constraint? 49
2.3.1. Conflations in the continuum hypothesis 51
2.3.2. Convention in usage-based grammar 57
2.3.3. A research agenda for understanding coselectional constraint 59

3. Hypotheses and data **62**
3.1. Hypotheses 62
3.2. Data 67
3.2.1. onDaF-based grouping 68
3.2.2. Annotations 72
3.3. Summary 77

4. Statistics **78**
4.1. Diversification 79
4.1.1. Lexical diversity 79
4.1.2. Coselectional diversity 82
4.1.3. Morphosyntactic diversity of verbs 83

4.1.4. Diversity in argument types . . . 85
4.1.5. Distribution of verb-argument structures (VAS) . . . 89
4.1.6. Summary: Diversity and Diversification . . . 91
4.2. Similarity . . . 91
4.2.1. Part-of-speech distributions . . . 91
4.2.2. Shared vocabulary . . . 91
4.2.3. Most frequent verbs and arguments . . . 96
4.2.4. Shared coselections . . . 105
4.3. Specialization and lexical association . . . 109
4.3.1. Statistical measures vs. the linguistic model . . . 110
4.3.2. Lexical association in Kobalt . . . 113
4.4. Summary . . . 135

5. A graph-based model of verb-argument coselection in Kobalt 137
5.1. Graphs as a data and knowledge structure . . . 137
5.1.1. Graphs and linguistic theory . . . 140
5.1.2. Summary . . . 144
5.2. The model . . . 144
5.2.1. Specifics of the model . . . 148
5.3. Graph structure and lexicosyntactic coselection . . . 163
5.4. Specified hypotheses . . . 169
5.5. Summary . . . 169

6. Results and validation of the graph-based analysis 171
6.1. Results by onDaF-group . . . 174
6.2. Graph specificity and subjects in L1 vs. L2 . . . 179
6.3. Validation . . . 186
6.3.1. Corpus size . . . 187
6.3.2. Grouping . . . 195
6.3.3. Summary: Corpus size and grouping . . . 217
6.3.4. Text length . . . 219
6.4. Summary . . . 230

7. Discussion and future research 233
7.1. Unexplained variance . . . 233
7.1.1. Typology . . . 233
7.1.2. Register, cultural, and teaching effects . . . 241
7.1.3. L1-standard, variance, and text-linguistic effects . . . 247
7.2. Methodology . . . 251
7.2.1. Replication . . . 251
7.2.2. Improvements to the linguistic model . . . 252
7.2.3. Larger data and sampling . . . 253
7.3. Graph-based modeling of linguistic phenomena . . . 256
7.3.1. The isograph problem . . . 257
7.3.2. Graph theory and network analysis in linguistics . . . 258
7.3.3. Grammar as graph . . . 260
7.4. Towards a theory of coselectional constraint? . . . 263
7.5. Summary . . . 268

A. Appendix **272**
A.1. Formal definition of the graph-based model 272
A.2. Approximation of the modularity limit 275
A.3. Weighted vs. unweighted modularity 276

Bibliography **277**

List of Figures

3.1. Histogram of the onDaF score distribution 71
3.2. Text length distribution vs. onDaF scores 72
3.3. Dependency parse of a sentence with a chain of coordinated verbs. 75
3.4. Dependency parse of the prompt 77

4.1. Type-Token-Ratio in individual documents in Kobalt 81
4.2. Transformed Type-Token-Ratio ($TTR \cdot \sqrt[4]{token}$) in individual documents in Kobalt 81
4.3. Ratio of unique V + dep combinations 83
4.4. Verb categories in Kobalt documents 85
4.5. Relative frequencies of verb dependents in individual documents 87
4.6. Relative frequencies of verb dependents in individual documents, free y-scale 88
4.7. VAS-distribution in Kobalt subcorpora 90
4.8. Distribution of POS categories in Kobalt by subcorpora and language . . . 92
4.9. Heatmap of lexeme overlap between texts in L1 93
4.10. Simplified heatmaps for lexical overlap between BEL texts 94
4.11. Simplified heatmaps for lexical overlap between CH texts 95
4.12. Mean, maximum and minimum lexical coverage through intersection by language group 96
4.13. 20 most frequent verbs in Kobalt 98
4.14. Percentage of 20 most frequent out of all verbs in Kobalt subcorpora 100
4.15. Absolute frequencies of 20 most frequent verbs in Kobalt subcorpora 101
4.16. Percentages of argument lexemes out of all lexemes, ranks 1-25 102
4.17. Percentages of argument lexemes out of all lexemes, ranks 26-50 103
4.18. Percentages of argument lexemes out of all lexemes, ranks 51-100 104
4.19. ΔP for V+OBJA coselection in L1, frequency ≥ 5 117
4.20. ΔP for V+OBJA coselection in L1, frequency ≥ 3 118
4.21. ΔP for V+OBJA coselection in BEL, frequency ≥ 5 120
4.22. ΔP for V+OBJA coselection in BEL, frequency ≥ 3 121
4.23. ΔP for V+OBJA coselection in CH, frequency ≥ 5 122
4.24. ΔP for V+OBJA coselection in CH, frequency ≥ 3 123
4.25. ΔP for V+SUBJ coselection in L1 124
4.26. ΔP for V+SUBJ coselection in L1 125
4.27. ΔP for V+SUBJ coselection in BEL 126
4.28. ΔP for V+SUBJ coselection in CH 127
4.29. ΔP for V+PRED coselection in L1 129
4.30. ΔP for V+PRED coselection in BEL 130
4.31. ΔP for V+PRED coselection in CH 131
4.32. ΔP for V+OBJP coselection in L1 132
4.33. ΔP for OBJP coselection in BEL 133

4.34. ΔP for OBJP coselection in BEL, BEL_075 does not have identical coselections for OBJP . . . 134

5.1. Example of a table-based co-occurrence visualization . . . 141
5.2. Dependency parse with several realized accusative objects (Foth 2006) . . . 150
5.3. Dependency parse with several realized accusative objects (new model) . . . 151
5.4. Dependency parse with two modal constructions (Foth 2006) . . . 151
5.5. Dependency parse with two modal constructions (new model) . . . 152
5.6. Example of a determiner connecting unconnected verbs and arguments in Kobalt L1 . . . 156
5.7. Example of a preposition in a prepositional object connecting unconnected verbs and arguments in Kobalt L1 . . . 157
5.8. Graph-based verb-argument coselection model of Kobalt L1 . . . 161
5.9. Graph-based verb-argument coselection model of Kobalt BEL-115 . . . 162
5.10. L1 no_subj graph and degree distribution . . . 164
5.11. BEL no_subj graph and degree distribution . . . 165
5.12. CH no_subj graph and degree distribution . . . 165
5.13. Complete graph of 10 nodes . . . 166
5.14. Empty graph of 10 nodes . . . 166
5.15. Barbell graph of 10 nodes . . . 168
5.16. Two random graphs of 10 nodes and 15 edges . . . 168

6.1. Modularity in onDaF-based grouping . . . 176
6.2. Variance of modularity in onDaF-based grouping . . . 177
6.3. Modularity in onDaF-based grouping, no_subj and vas_no_prep only . . . 178
6.4. Modularity in 10-text samples from onDaF-based subcorpora, scatterplot . 180
6.5. Number of unique subject and object lexemes in documents . . . 181
6.6. Unique OBJA and SUBJ lexemes in individual documents . . . 182
6.7. Ratio of unique OBJA/unique SUBJ lexemes in individual documents by text length . . . 183
6.8. Frequent subjects in L1 but not L2 . . . 184
6.9. Frequent subjects in L1 but not L2 . . . 185
6.10. Modularity of graphs derived from individual documents . . . 189
6.11. Modularity of no_subj and vas_no_prep graphs derived from individual documents . . . 190
6.12. Modularity of no_subj and vas_no_prep graphs derived from individual documents vs. text length . . . 191
6.13. Number of documents in onDaF10-based subcorpora. . . . 192
6.14. Modularity for onDaF10-based subcorpora, five 6-text-samples per box . . . 193
6.15. Modularity in onDaF10-based subcorpora, five 6-text-samples per box . . . 194
6.16. Modularity in 10-text samples from earlier onDaF grouping, 5 samples per subcorpus, reproduced here for comparison, identical to fig. 6.3 . . . 194
6.17. Comparison of 5/6- and 9/10-sampling in L1 . . . 196
6.18. Comparison of 5/6- vs. 9/10-sampling in BEL, vas_no_prep . . . 197
6.19. Comparison of 5/6- and 9/10-sampling in CH, vas_no_prep . . . 198
6.20. Comparison of 5/6- and 9/10-sampling in BEL no_subj . . . 199
6.21. Comparison of 5/6- and 9/10-sampling in CH no_subj . . . 200
6.22. Variance of modularity in 5/6- and 9/10-samples . . . 201
6.23. onDaF score ranges in sliding windows . . . 203

6.24. Modularity vs. sample size . . . 204
6.25. Median modularity vs. sample size . . . 205
6.26. onDaF median vs. onDaF mean in sliding windows . . . 205
6.27. Modularity median in corpus sizes by onDaF groups and language, vas_no_prep only . . . 206
6.28. Modularity median in corpus sizes by onDaF groups and language, no_subj 207
6.29. Sliding windows compared . . . 208
6.30. Sliding windows compared with automatically fitted regression line . . . 210
6.31. Sliding windows of 5 texts compared . . . 211
6.32. Sliding windows of 10 texts compared . . . 211
6.33. Sliding windows of 15 texts compared . . . 212
6.34. Sliding windows of 20 texts compared . . . 212
6.35. Sliding windows compared (BEL) . . . 213
6.36. Sliding windows compared (CH) . . . 214
6.37. Sliding windows compared (L1) . . . 215
6.38. Text length and onDaF medians in windows (L1) . . . 216
6.39. Text length distribution by onDaF-scores . . . 223
6.40. Modularity vs. text length median in 15-text-windows (L2) . . . 225
6.41. Modularity vs. text length median in 10- and 15-text-windows (L1) . . . 225
6.42. Modularity in 20-text-windows based on 450+ tokens and full texts, free y-scale. . . . 226
6.43. Modularity in 20-text-windows based on 450+ tokens and full texts, free y-scale, with approximate trajectory . . . 227
6.44. Number of VAS in individual texts . . . 228
6.45. Modularity in first vs. last 40 VAS in sliding windows (15 texts) . . . 228
6.46. Modularity in first vs. last 40 VAS in sliding windows (15 texts), free y-scale 229

7.1. Modularity by text length and onDaF median in L1 . . . 248
7.2. A graph model of lexicosyntax in Kobalt L1 . . . 261

A.1. Approximation of the modularity limit . . . 275
A.2. Weighted vs. unweighted modularity in onDaF-based grouping . . . 276

List of Tables

3.1. Number of documents in Kobalt subcorpora 71
3.2. Number of tokens in Kobalt subcorpora 71
3.3. Prompt parsed into heads and dependents 77

4.1. Unique verb lexemes in Kobalt subcorpora 80
4.2. Unique argument lexemes in Kobalt subcorpora 80
4.3. Unique verb + dependency type coselections in Kobalt 82
4.4. The 12 coselections that occur in all ten subcorpora of Kobalt and absolute frequencies. 106
4.5. All coselections that occur in all BEL subcorpora of Kobalt and absolute frequencies. 107
4.6. All coselections that occur in all CH subcorpora of Kobalt and absolute frequencies. 108
4.7. Contingency table for factor combinations such as word co-occurrences . . . 114
4.8. Contingency table for the coselection of *haben + Problem* ('to have + problem') in Kobalt L1 114

5.1. Example of a table-based co-occurrence visualization 141

Acknowledgements

Like anything worth doing, this work is the result of a communal effort and a synthesis of inspiration drawn from many kind and resourceful people.

First and foremost, I would like to express my gratitude to my supervisors, Anke Lüdeling at Humboldt University of Berlin and Amir Zeldes at Georgetown University. Everything I know about modeling and questioning data, and relating models to models rather than questions to answers, I have learned from Anke. The environment that she has created at HU Corpus Linguistics is communicative, supportive, and critical in the best sense. It has given me a mass of opportunities and a home, for which I am deeply grateful. Amir took on my work mid-project and I have profited immensely from his talent to give pointers to just the right reference for any question, and from his detailed, knowledgeable, and critical feedback on my writing. I bow to his quick wit and his sharp and informed reasoning, and hope to become a worthy sparring partner in future discussions. I would also like to thank Hans Boas at the University of Texas at Austin, who was a very kind and supportive advisor during the early stages of this work, and who made a research stay in Austin possible for me during which the complications of the topic became apparent and thus eventually available for disentanglement.

This work was graciously funded by the Hans Böckler Foundation through a BMBF scholarship in the years 2014–2018, and by the Humboldt Graduate School through a Research Track scholarship that allowed me to collect ideas for my research proposal in 2013. I would like to thank all those who were involved in raising and distributing these funds for this opportunity. I would further like to thank Prof. Dr. Christoph Möllers and my colleagues at the Humboldt University Faculty of Law for giving me the time and space to finish this project.

The study would have been entirely impossible without the Kobalt team, who collected the data in four countries and generously shared it with me, for which I would like to thank everyone involved. It would have been equally impossible without the support of many people at the Department of German Studies and Linguistics, Eva Schlachter and Birgit Trettin in particular; and the colleagues at HU Corpus Linguistics, who were always supportive and generous with their time and expertise. Everyone on the team is wonderful and has shaped this work, but special thanks are due to Felix Golcher, whose doubts in lexical statistics have given me much clarity, and who has been of immeasurable help with many things involving R; to Hagen Hirschmann, whose linguistic insight I value much; to Konstantin Schulz, whose linguistic and methodological comments are always helpful, as was his support with JavaScript; and in particular to Thomas Krause, who has helped me greatly with the computations in this work and many other engineering things, who is an excellently skeptical colleague, a valued partner in discussions of scientific methodology and reasoning, and a great friend.

I would further like to thank Katrin Wisniewski, Vivien Mast, Annarita Liano, and Margo Blevins for helpful and always smart comments on my writing, for proofreading, and for being exceptionally amazing in all regards. Special thanks are also due to Matthias Fingerhuth, who encouraged me to get back on the horse when things did not seem to work out at all, and who sparked my interest in the epistemology of corpus linguistics. Unbeknownst to her, my aunt Natalia Alexandrovna Shadrova is involved in this work with a remark that stayed with me from when I was a teenage heritage speaker of Russian stumbling through the jungles of words that go together and those that don't. I thank her for this.

I am grateful also to a number of teachers: Hans-Hermann Penning and Karl-Heinz Pitz, who taught me that math is beautiful and strange, and nothing to be afraid of; Katja Cantone-Altıntaş, who was a wonderful supervisor during my undergrad studies and who acquainted me with the wonders of empirical work; and Tomi Washida, who taught me to be aware of all the details, and that making new mistakes counts as progress.

Finally, my gratitude goes out to my grandmother Galina Dmitrievna Shadrova, who has raised many strong women and has always believed in me; to my grandfather Alexandr Nikolayevich Shadrov, who was gifted with a sharp mind, a loving heart, and skillful hands, and who saw me; to my father Valeriy Chikarev, for the music and the open arms; and to my mother Elena Yost, who still pulled us through. I owe my deepest thanks to the ancestors, for creating and tending to this world, for their gentle guidance and their blessing, and to the benevolent spirits for their wonders and their gifts.

And most of all I thank Luke, who with his soft and loving kindness and his quirky ways is my fluffy inspiration and my one true love.

1. On fixedness, conventionality, and coselection

> "Since language is conventionalized by its very nature, the term conventionalized language in this paper is used to mean "language forms that are *more* conventionalized than other language forms"" (Yorio, 1989, 56, emphasis in the original).

Aspects of 'uncreative', repetitive, or static ways of language use were noticed early on in the history of linguistic study. Holistic, non-spontaneously created speech was first reported in patholinguistics in "the view that some language was 'automatic' and 'non-propositional'" in patients with brain damage (Wray (2013, 320) in reference to Hughlings Jackson (1874)) and much research since has shown that indeed, fixed phrases are available in patients with aphasia and dementia who have otherwise lost their generative linguistic abilities.[1] This suggests that language is generally not one holistic system, but can be subdivided into interrelated processes, one group of which is able to handle information that cannot anymore, or not yet, be processed otherwise. Sequences of fixed elements, incongruent with mature native speaker grammar, but communicative in function, were also observed in young children acquiring their first language, and in learners (Braine, 1963; Pfaff et al., 1980; Titone, 1969). This is how fixed speech was studied first as a deviation from the norm, thus sometimes judged as peripheral to the 'actual' language of an unimpaired adult native speaker. But if the human linguistic system is subdivided into more automatic and more generative subprocesses, it should be divided in this way in all speakers, and speakers with a regular speech capacity would be expected to also make use of both. And indeed it has been shown in a series of papers that formulaic sequences are recognized, processed, and rated faster in learners and native speakers[2] and Siyanova-Chanturia (2013, 262) concludes in her review of eye-tracking and neurolinguistic, event-related potential (ERP) studies

> "that the processing of MWEs [multi-word expressions, AS] differs from novel language not only quantitatively — that is, in terms of the speed of processing; but also qualitatively — that is, fundamentally distinct neural correlates underlie on-line processing of novel language and MWEs".

Chunks – continuous and fixed linguistic units that can go morphosyntactically unanalyzed – are not only processed faster, they are also important for fluent speech in native speakers and in learners, where it is not only the correctness of the chunk itself, but also its

[1] See for example Van Lancker (1988); Wray (2002); Maclagan et al. (2008); Leyton et al. (2014); Lindholm and Wray (2011); Sidtis et al. (2009); Davis and Maclagan (2010).

[2] Faster reaction times and lower error rates for non-compositional idioms (Jiang and Nekrasova, 2007) and compositional, but frequent 4-grams (Arnon and Snider, 2010); Shorter fixation times for final words of formulaic sequences in eye-tracking experiments (Underwood et al., 2004) and shorter reading times for idioms (with non-compositional meaning) vs. novel phrases in native speakers (Conklin and Schmitt, 2008; Siyanova-Chanturia et al., 2011); Shorter phonetic duration for frequent multi-word units even across syntactic boundaries (Arnon and Cohen Priva, 2013).

holistic pronunciation, i.e. the positioning of pauses at chunk boundaries, that is relevant for perceived fluency.[3] This is of course the case in language impairments, where fluency can *only* be maintained through chunks, but there are much more impressive examples of memorization in the service of fluency in unimpaired speakers. Chunks of enormous length are memorized in oral tradition, in religious recital and ritual, or in theatrical plays, like Greek tragedies with extensive monologue (see Pawley et al. (2007) for an overview of the study of chunks outside of linguistics). A productive reconstruction of such chunks would certainly be much less eloquent and detailed, but it would also be much less fluent.

With these observations, a general acceptance of fixedness as a fundamental aspect of natural language has come about in recent decades. To give but three examples out of many:

> (1) "It is widely accepted that natural language is to a great extent made up of multi-word expressions or formulae" (Valsecchi et al., 2014, 1);
>
> (2) "It is becoming increasingly apparent that language is largely formulaic in nature, and that the competent use of formulaic sequences is an important part of fluent and natural language use (...)" (Durrant and Schmitt, 2009, 157);
>
> (3) "(...) it is not necessary to study large corpora to show that everyday speech and writing depends heavily on conventional expressions. Analysis of quite small samples of spoken or written text are sufficient to show this" (Pawley et al., 2007, 22).

Estimations to the extent of chunkiness in natural language, particularly in spoken language, reach from 30-40% (Saito, to appear) over 55% (Erman and Warren, 2000)[4] and as high as 70% (Altenberg, 1991). They can be even higher for lexicosyntactic niches: Calude (2008) reports 90% of demonstrative cleft constructions in the Corpus of Spoken New Zealand English (200 000 token) to fit into the same schema: [DEMONSTRATIVE PRONOUN + BE + WH-WORD + RELATIVE CLAUSE], like *That's what he thought*, where each of the elements of the construction is a chunk. If one allows for overlapping chunks (*That's what + what he + he thought*), some would argue that up to all language in use is made

[3] For a recent overview of studies into the interrelation of formulaic language and fluency, see Tavakoli and Uchihara (to appear) and Guz (2017).

[4] It should be noted that, while the paper is widely cited and was one of the first to attempt a quantification of chunks at all, the way Erman and Warren count what they call *prefabs* is not highly transparent. It resembles an n-gram approach where in each n-gram, a slot can be filled with either a word or a prefab. Prefabs for them are any two or more words that occur together habitually (like *I think*) or functionally (like *instead of*). However, what constitutes a prefab and what does not is not entirely clear from their paper. They name as their main criterion *restricted exchangeability*, i.e. non-compositionality and syntactic fixedness, or failure to exchange one of the words without the phrase becoming unidiomatic, but then define "I went to some seminars" as a prefab without further explanation (ex. 1, p. 34). One is left to presume that they expect a native speaker to be unable to find a similarly idiomatic sounding way of expressing that thought. This is a problematic and circular argument prone to confirmation bias, because if I assume that specific meaning can only be expressed in a certain way, then I will not allow part of it to be exchanged because the meaning would then change – meaning I derive fixedness from use and use from fixedness. Such a rater judgment may also be an effect of prototypicality or simply priming, where, once a structure is primed and activated, it appears as the most natural. Under this restriction, 55% of familiar and idiomatic-sounding structures appears much less impressive.

up merely from chunks and chunks with slots.[5]

All of those observations refer to fully or nearly fully fixed language material like chunks, idioms,[6] and formulaic sequences (see Schmitt (2004); Wray (2002, 2012, 2013); Wood (2015) for more comprehensive research synthesis). But there exists also another kind of 'uncreative' or conventional material that is not 'automatic' and 'non-propositional', but still not freely combined. This what the quote from the beginning of this chapter points at: Out of many grammatical and contextually comprehensible ways to express meaning that are available through grammatical generativity, there appears to exist a much more limited number of expressions that seem natural or idiomatic in natural language. With Pawley and Syder (1983), this creates two puzzles for linguistic theory: Native-like fluency and native-like selection. And when asked why an unidiomatic combination does not work, native speakers are often lost for an explanation beyond "yes, you could say that…But you *wouldn't*".

One of the first to point this out at word level was Firth (1957, 11) in his seminal quote "you shall know a word by the company it keeps", where he refers to collocations – words that tend to habitually co-occur, without necessarily being a phonetic or graphematic chunk – and to different word senses that lexemes obtain in different contexts. Such forces of attraction that lead to a thematic clustering of words may be semantic, conventional, or otherwise.[7] They appear to exist not only on a surface and word level, but also in the preferences of lexemes to occur in certain syntactic environments over others, in the covarying co-occurrence of two lexemes in two slots of a construction, in the distribution of verbs on verb-argument structures or subcategorization frames, as well as in diverging lexical choices in seemingly synonymous syntactic alternations. This will be discussed in some detail in section 2.1. Preferences of this kind appear to be at least partially idiosycratic in that they cannot be predicted from semantics or syntax alone, and to exhibit distributional properties unique to one of the lexemes they are bound to. Sinclair (1991) famously coined this the *idiom principle* as opposed to the *open choice principle*:

> "[T]he principle of idiom is that a language user has available to him a large number of semipreconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991, 110).

It is this kind of conventionality that is of interest in this thesis, and items that co-occur in this way will therefore be referred to as coselections or coselected items.

The notion of coselection can be read in one of two ways: Either two items are coselected from a lexical set simultaneously; or one item coselects another. This aspect will be further discussed where the data suggests one thing or another, and in the discussion in chapter

[5] As is done in some more radical lexicalist grammar models, like pattern grammar (Hunston, 2012) and word grammar (Hudson and Hudson, 2007). This raises problems of identification for an empirical verification though. Since words in a text are not randomly but syntactically arranged, most bigrams that include one of the words of the presumed chunk will also be limited to a number of lexemes that can syntactically be adjacent to the word; and those will be distributed according to lexical frequency distributions. In other words, if the words 'he' and 'thinks' re-occur in speech, and I find that they also co-occur in a reference corpus, that is not direct evidence of a chunk but of the adherence to the same distribution, i.e. syntactic rules.

[6] Idioms are usually defined as relatively, but not entirely fixed units with non-compositional meaning, such as *to spill the beans*.

[7] More will be said about this conceptualization in section 2.3.

7. Whether or not conventionality affects all multi-word units, chunks, coselections, etc. similarly, will be discussed in section 2.3.1.

Unlike chunks, which can take complex syntactic forms but appear to be treated like words in processing and production and thus are essentially lexical items, coselections always involve a syntactic context. Not only are a verb and and its argument lexeme coselected, but the argument is also necessarily coselected with a specific syntactic slot – it cannot be realized outside of a syntactic context, like accusative or prepositional object slot – and both are coselected with a syntactic construction like one variant of an alternation. If syntactic structures exert forces of attraction on specific lexemes, then one of the questions is how this interacts with the collocational association between two lexemes. Coselections are therefore at the very meeting point of lexis and grammar: A lexicosyntactic phenomenon.

Building on the observation that on the level of *langue*[8] as represented in corpora,[9] language seems to be to some degree arbitrarily, idiosyncratically, or distributionally constrained, and that these constraints cannot be easily predicted and hence must be learned in L1 and L2, this thesis is interested in the use and development of verb + argument coselections in native speakers and learners of German, and more specifically, in the development of coselectional constraint. Coselectional constraint is a conceptualization of the limitations of combinatorics, i.e. the force or principle that keeps words from entering combinations freely. It is both a feature of each word and a feature of the language as a whole. This will be discussed in detail in chapters 4–6.

An in-depth study of the development of coselectional constraint may be interesting for several reasons:

A number of studies have shown the constraining tendencies listed above, but those all project to the *langue* in being performed on larger corpora that do not allow for an estimation of inter-individual and stratified variance; and they are typically not hypothesis-based beyond the hypothesis that distributions will differ by lexemes or syntactic constructions. For an overview of these, see section 2.1. Little is in fact known about the extent of coselectional constraint as such, it has only been observed in terms of variable distributions of individual constructions. To date, there is no hypothesis-based, contrastive study into the degree of coselectional constraint across verbs, and how it differs between learners and native speakers.

Secondly, speakers show distributional sensitivity in learning new words or constructions, and several models of language learning assign structural importance to repetition, segmentation and analysis in FLA and SLA. But the degree of importance and the role of *conventionalized* rather than *fixed* structures is at present not well-researched: While fixed units can only be used in fixed ways, coselectional preferences can be conceptualized as constraints on the set of candidates for arguments or collocates. These are not fixed

[8] Language of a speech community as opposed to that of an individual, *parole* (Saussure, 1916/1983). Some argue that idiomaticity or conventionalized language exist only in the speakers' minds (Wulff, 2008). But convention can only be acquired through exposure, it only emerges in a group (otherwise, it would not be convention, but individual preference), it is particularly useful in a group, because it limits the semantic and semoiotic search space in interaction (Wray, 2002), and due to long-tailed lexical distributions, it is realized much more clearly in the *langue*, otherwise there would be no need for corpus-based collocation extraction. More on this in section 2.1.1.

[9] Whether or not this is a good representation will be discussed further in chapters 4 and 6.

in many ways, for example they still require syntactic embedding, and the idea of a constraint (rather than a forced choice) implies a degree of flexibility. Constraints can exist in two ways: Distributionally, in how many or what general type of candidates are allowed or likely, and exemplar-based, in which candidates specifically are chosen. A syntactic structure may prefer one item specifically out of a set of many others, or it may prefer one item slightly over only two more. Both are constraints on the coselectional set, but with different structural, combinatorial, and quantitative repercussions. In SLA, coselection is mostly studied in the shape of collocations, and most research is interested in the over- and underuse of collocational exemplars or types of collocations compared to an L1 standard, often for language assessment purposes. However, there appears to be no research into the structural development of coselectional preferences in learners, and research into the use of collocations and what is sometimes called phraseological complexity is not typically discussed as part of a systematic model in one of the frameworks of interlanguage (Selinker, 1972) or language variety (Klein, 1998; Klein and Perdue, 1997). Even where theoretical frameworks like emergentist or connectionist models are presumed, results are not discussed within such frameworks to come to a more complete and more fine-grained theoretical model of SLA, and no models contrasting conventionality with fixedness and their structural roles in L1 or L2 exist to my knowledge either. This will be discussed in section 2.2.

And thirdly, at this point, usage-based linguistics has emphasized the role of idiosyncrasy, specification, frequency, and distribution in its model of natural language. But there is vagueness as to the role of convention and idiosyncrasy in general and discrepancies and imprecisions regarding a number of concepts. Convention and idiosyncrasy are often equated and also confounded with one of the concepts of frequency or fixedness (or both), or used as an external cause – and a circular argument – for the way language works (we use language in this way because it is conventional, and it becomes conventional because we use it in this way). This is manifested in the model of a monodimensional continuum from fixed to freely combined language, where coselections of the kind discussed here are situated somewhere in the middle. There are, however, good reasons to consider a multidimensional model of recurrent use or coselectional constraint, where convention and fixedness are only some of the dimensions; and from the history of linguistics it is also resonable to assume that what keeps reappearing on a number of levels may have a function in its own right. This will be discussed in section 2.3. The chapter will conclude with a research agenda for the corpus study that is performed in chapters 4–6 (more will be said about this shortly). Concretizations of the research question and hypotheses can be found in chapters 3.1 and 5.4.

It should be noted that, while much of the referred literature is focussed on collocations, coselectional constraint and collocations do not in fact denote the same concept. A collocation is an exemplar and an expression of the coselectional constraint of both collocates. A set of collocations as they are usually studied includes a number of constrained coselections, but not typically of the same words (collocations are usually studied categorially, like adverb + adjective collocations, not paradigmatically for a lexeme), and it does not look into uses of each of the collocates with other, less restricted words. In other words, collocation studies are interested in a comparison of how many and which restricted combinations of words are used, while this study will look into how many and which arguments verbs take, restricted or unrestricted. If a word only ever occurs with

a single other word, then the collocation reflects a strong coselectional constraint. But it will be shown in chapter 4 that often, the same verbs are used with lexicalized coselections (idioms, support verb constructions, etc.) and productively with new arguments. While these uses may differ in salience and frequency, they are still both entrenched with the same verb. This means that in order to learn both the productive, generalizable aspects of a verb, and the specialized constraints in certain contexts, an organization of the coselectional distribution – including their constraints, inner distributional properties (Zipf or otherwise), exemplars and prototypes, and productivity – of each lexicosyntactic coselection is necessary (Zeldes, 2012, chapter 6). The central thesis of this study is that the acquisition of coselectional constraint is a structural process of lexical (re-)organization and lexicosyntactic differentiation, rather than a gradual increase in collocational exemplar knowledge.

It is, however, one thing to conceptualize a reorganization of lexicosyntax in a theoretical account, and a rather different thing to model and operationalize concepts in a way that allows for a quantitative analysis. While there is some methodology for measuring productivity (see Zeldes (2012); Baayen (2001) for an overview with many references), the methodological landscape regarding coselections is somewhat scarce. What at the beginning of the writing process of this thesis seemed like a simple task – applying statistical measures as they have been discussed for 15 years to existing data from a homogeneous corpus of essays written by native speakers and learners of German at different stages of acquisition – turned out to carry methodological difficulties that have not yet found much attention in usage-based linguistics.[10] There are at least three general problems with statistical analyses of coselections:

1. Statistics measure against randomness or against similarity (are two models divergent from being fully random or from one another?). But with the extreme combinatorial power of word distributions, randomness is a very poor baseline; and without a quantitative model of the *idiom principle*, it is unclear what would constitute a *sufficiently* non-random result. See section 4.2.4 in particular for a detailed review of the combinatorial power of even the smallish corpus used, and the problems arising from this in a quantitative analysis.

2. A statistical analysis relies on the comparison of identical factors, like lexemes, across groups. While it is possible to account for the presence or absence of items, it is impossible to define the coselectional constraint of an absent item in a group. Thus, lexical diversity between groups leads to a lack of comparable items. In addition, structural information cannot only be derived from individual items. The definition of a system is the set of its parts *and their interrelations* (Mesarovic, 1964). If lexicosyntax is a system, looking into individual factor combinations will not illuminate the full scope of implied processes.

3. There are deep epistemological problems with a number of assumptions in statistical analyses, like assumptions of randomness or independence of items (Schmid, 2010; Kilgarriff, 2005; Koplenig, 2017) and with the ontological status of the corpus as a presumed sample from a superpopulation, i.e. a probability-bound (stochastic) system that defines corpus outcomes. This is even more problematic in sparse, smaller-sized, and more variable data, like it is typical for the study of non-canonical

[10] See chapter 3 for a detailed description of the data.

language. A focus on statistical analyses does not satisfy the need for exact methods and measurability in non-canonical or sparse data; and a reliance on large data leaves virtually all analyses linguistically underspecified compared to the full analytical capabilities of the field. Relying on large data alone will always mean leaving out the many levels of linguistic analysis that require manual annotation due to structural ambiguity or due to low parser and tagger preformance; but also those that are simply not easily detected by machines, like rhetorical structures.[11]

This is why the ultimate focus of this study is largely methodological. Chapter 4 provides a statistical analysis of lexicosyntactic aspects of the data, the deeply annotated and manually corrected corpus Kobalt (Zinsmeister et al., 2012). It consists of 151 essays written by learners of German from Belarus and China, and 20 control texts written by native speakers. The analysis in a number of measures points at a process of reorganization and structural changes in coselectional preferences and constraints of verbs and nouns. At the same time, it also shows that specific, exemplar-based coselectional constraints are rather difficult to quantify in a linguistically meaningful and interpretable way even where lexical overlap is large, i.e. texts appear 'similar enough' for comparison. While certain conventionalities and idiosyncrasies are visible, only a vague understanding emerges even with the triangulation of a number of statistical measures. Also, while the observation of conventionality in natural language has been persistent on a *langue*-level, quantifying coselectional preferences on a *parole*-level, or cross-sectionally for subsets of the *langue*, is hindered by the sheer unfolding of the combinatorial power of words. Texts, even those that are written on the same topic by a similar cohort, are similar in many ways, but there is not a huge number of identical coselections across cohorts. To what degree this is a matter of data size will be discussed in chapter 7 – other explanations are that the types of coselected items change and evolve in learners, that learners and native speakers choose different registers and styles, and that communicative expression simply differs inter-individually, and even intra-individually in different text parts.

To gain a clearer understanding of the *structural* changes involved, a structural, namely a graph-based, model of lexicosyntactic coselection is developed in chapter 5. Graphs are used widely for the visualization of linguistic analyses (most notably as syntax trees) and more recently in the wider digital humanities, but graph metrics have not yet been applied much in linguistics. They offer a structural perspective, in which the relations between all elements can be considered simultaneously, rather than additively through the separate analysis of item-based factor combinations, and are used in all STEM fields in a variety of research questions around complexly interacting systems. Their application on natural language informed by linguistic understanding looks promising for a new perspective on structural developments. This will be further discussed in chapter 7.

But first, a metric of the internal structuredness of a graph called *Louvain modularity* (Blondel et al., 2008) is applied to cross-sectional subcorpora in chapter 6, showing in surprising clarity that indeed, lexicosyntactic organization undergoes structural changes, and that usage-based grammar and interlanguage theory have predictive power if applied in a hypothesis-based manner. It also shows an unpredicted difference between the two learner cohorts, where a process of differentiation and specialization appears to be of a more linear quality in Chinese learners, while Belarusian learners appear to show a later, longer, and much stronger period of randomization, leading to an overall u-shaped development in the

[11] For an analysis of rhetorical structures based on manual annotation on the same corpus used in this study, see Wan (in prep.).

latter, but not the former. This divergence may be an artifact of the data, but it is well possible that other linguistic aspects – differences in typology, language environment, and textuality – can provide an explanation for it. This will be discussed in chapter 7.

With this graph-based quantification, the study provides a first application of a new method for assessing structural aspects of lexicosyntax even in a smaller corpus like Kobalt. Specifically, it shows that Louvain modularity on this type of text in this register and the cohorts analyzed converges after less than 50 texts in L2, and less than 30 texts in L1. If this can be confirmed in future work, it might be a contribution to the development of exact and quantitative models of linguistic subfields that are intrinsically limited to working with sparse data. However, replication and extension are necessary first, aspects of which will be discussed in chapter 7.

The study also aims at contributing to the development of a methodology for the development of methods in corpus linguistics (recursion intended), putting an emphasis on a detailed validation against confounding factors such as text length or corpus size, and a critical discussion of results in chapters 6 and 7. Suggestions to the potential for an extension of the method to other research questions are also made in chapter 7.

It will be shown that careful theorizing, modeling, and validation allow for a *more linguistic* perspective on the data than is often the case in quantitative corpus studies, and that this in turn yields interesting results that raise further questions to the role of conventionalized or constrained coselection. This also exemplarily suggests that corpus linguistics has the potential of developing a richer methodology capable of capturing complex effects without making many of the concessions required by methodological premises today. If it did, this would mark a step forward not only in the methodological, but also the epistemological development of the field.

Finally, results are tied back to the linguistic background discussed in chapter 2, closing with a suggestion of a less frequentist and more functional view of coselectional constraint. Being located not only at the center of lexicosyntax, but also of *langue* and *parole*, the issue will show itself as a highly complex phenomenon intertwined with aspects on various, and perhaps all, linguistic levels. A full disentanglement of all complexities will not be possible within the scope of this thesis, but a first approximation is aspired to in the guiding questions of (1) how does coselectional constraint develop in learners, and how does it manifest in the writing of native speakers?; and (2) how can this be captured empirically without greater linguistic concessions to the quantitative model? In answering this, the thesis provides a first step towards viewing the *idiom principle* not only as a metaphor or a generalized feature of natural language, but as a measurable aspect of speaker and community language that varies in different cohorts as well as individually.

Indeed, it seems clear that some language forms are *more* conventionalized than others (see Yorio (1989) from the beginning of this chapter). But beyond this birds-eye-view description, the details and text-linguistic entanglements in this are not only fascinating, but seem to hold much valuable linguistic insight, once the challenge of their empirical measurement is overcome.

2. Linguistic background

This chapter provides an overview of the linguistic background of coselectional preferences or conventionalized co-occurrence of linguistic items. It first presents observations regarding the role and use of coselections in corpus-based studies in section 2.1, showing that while coselectional constraint has been observed on a number of levels, its specificity, stability and extent are currently unknown, particularly as it it is represented in individual speakers or in speaker cohorts. Section 2.2 discusses the role of convention and coselection in language learning with a focus on learner corpus research, distributional sensitivity in L1 and L2, and (the lack of) an integration into systematic models of SLA. The chapter concludes with the observation that while corpus studies and models of language learning point towards differences between fixedness and convention in many ways, the two are still modeled on a monodimensional continuum, which holds discrepancies and conceptual imprecisions. Those are discussed in section 2.3, closing with a research agenda for a view of coselectional preferences and constraints in their own right.

This review will be limited in scope, and more can certainly be said about the issues discussed. An in-depth research synthesis of all concepts related to coselectional constraint and conventionalization in language as they exist in corpus studies, in language learning, and teaching is desirable; and in particular, a detailed analysis of the concepts that imply certain things about conventionality, often latently manifested, in the theories of usage-based grammar is required for an eventual model of coselectional constraint. It cannot be provided within the scope of this thesis though, for one thing because, being located at the center of lexicogrammar and thus related to several extensive fields of research, the task and the material suffices for another book. But mostly, because the focus of this thesis is largely a methodological one, and the proposed method itself also requires synthesis from a number of fields. Thus, more literature will be reviewed in the other chapters. This chapter can therefore only report on the main strands of research, but not all details of the related discussions. Its major goal is two-fold: To report where and how coselectional constraint has been observed, and to discuss how it is not yet integrated into existing usage-based models in a clear way, and why more research with a focus on coselectional constraint is therefore needed.

2.1. Coselection in corpus studies

Coselectional constraint on a word level is generally studied under the umbrella term of *collocations*. This research is reviewed with a focus on native speakers in section 2.1.1. Constraint in the coselection of words and syntactic units has been observed in two main lines of corpus research: One is distributional and largely statistical, often termed *collostructional analysis*. The other is more qualitative and classificatory, and focuses on the idiosyncrasies of words. Although there are no clear boundaries and some studies combine both approaches, this is how the remainder of the section is divided: Collostructional distributions (section 2.1.2), and idiosyncrasy (section 2.1.3).

2.1.1. Collocations

While conceptually, collocations as such are not at the heart of this study, they are the most studied coselections in applied linguistics, specifically in mono- and bilingual lexicography[1] and L2 teaching[2]. There is another school of collocational research inspired by Eastern European, functional traditions, see for example Mel'cuk (1996); L'homme and Bertrand (2000); Bartsch (2004), where collocations are explicitly modeled as coselectional processes, i.e. one collocate coselects the other. However, these approaches are not typically empirically oriented in the sense that they would be interested in speaker cohorts, but have mainly lexicographic and classificatory aims, similar also to the work of Sinclair (1991, 1996, 1999). Since this thesis is concerned not as much with aspects of the *langue*, but with developmental processes and stratified variation, those will not be discussed here further. It is important to note that an empirical approach does not imply a focus on stratified variation. Corpus-based lexicography very much is empirically oriented, but is still interested in the collocations per se, not in what shapes their use.

More will be said about the acquisition and use of collocations in learners in section 2.2. However, since research into collocations in use that takes into account individual speakers is typically realized in contrastive L1-L2 studies, some overlap cannot be avoided, although statements to the form, extent, or function of collocations in L1 are typically only implicitly made:

> "(...) there are several issues that compound the difficulty of acquiring L2 collocational knowledge and these include, to name just a few, a lack of perceptual salience and deceptive transparency of many MWUs [multi-word units, AS], cross-linguistic variability of collocational forms (e.g., delecixalized phrases such as *make a mistake*), irregular spacing of encounters with phrases, and a traditional focus on teaching indiviual words rather than MWUs" (Szudarski, 2017, 206).

This description is formulated in reference to the teaching and acquisition of collocations in a classroom setting. But a lack of perceptual salience, the existence of coselectional

[1] Some of the central questions are how to structure a collocational lexicon; what to include based on statistical measures or linguistic features; how to map collocations in a bilingual lexicon, see Benson (1989); Kilgarriff and Tugwell (2001); Evert et al. (2017); Bouma (2009); Pecina (2010); Evert and Kermes (2003); Krenn et al. (2001); Dias et al. (2000); Almela (2011); El Maarouf et al. (2014); Evert (2008); Malmkjaer (1993); Barfield and Gyllstad (2009); Brooke et al. (2015); Steyer (1998, and others); and how they can be used in machine translation, Koehn (2005); Cohn and Lapata (2007); Ohmori and Higashida (1999); Orliac and Dillinger (2003); Liu et al. (2010); Pearce (2001); Seretan and Wehrli (2007); Seretan (2011); Uhrig et al. (2018), and, similarly, for word sense disambiguation in natural language processing, see Schneider (2014), El Maarouf et al. (2014).

[2] The literature on language teaching and languages for specific or academic purposes, of which collocations are an essential interest, cannot be reviewed here. For an overview of collocations in language teaching with a number of references, see Targońska (2019); Wray (2013); Meunier and Granger (2008); Szudarski (2017). A few references to intervention studies in language teaching and teaching methods: Barfield and Gyllstad (2009); Vassiljev et al. (2015); Kennedy (2003); Boers et al. (2014a); Lindstromberg et al. (2016); Szudarski and Carter (2016). For some work on English for academic and specific purposes (EAP/ESP), and English as a lingua franca (ELF), see Simpson-Vlach and Ellis (2010); Nagy and Townsend (2012); Wood and Appel (2014); Granger (2017); Guerrero (2004); Hyland (2006); Rausch (2016); Hancioğlu et al. (2008); Kecskes (2007). There is also limited work on other languages for academic and specific purposes, like French (Martin, 2010; Beeching, 1997; Noe, 2003; Ryabova and Sergeychick, 2018; Owoeye, 2010), Spanish (Mendoza and Knoch, 2018; Sánchez-López, 2018; López, 2015; Doyle, 2018), Mandarin Chinese (Quan, 2011; Tao and Chen, 2019), and German (Wallner, 2014; Kärchner-Ober et al., 2015; Jaworska, 2015).

constraint in spite of the lack of syntactic and semantic idiosyncrasy, as well as high variability regarding the frequency of occurrence make collocations an overall strange phenomenon that is not easy to grasp in a linguistic model. In addition, there are two majorly confusing aspects in the discussion of collocations:
Firstly, there is a variety of terms in the literature where it is often not quite clear what is meant exactly. Wray (2002) lists 56 different terms ranging from *formulaic sequences* to *lexical bundles* to *n-grams* or *collostructions*, and several more have been used since. All of these terms may, but do not necessarily, refer to collocations. There is also a number of somewhat disconnected research subfields around some of the terms which leads to an overwhelming number of related studies that are not always easy to find; and while most define whether they refer to fixed, idiomatic (non-compositional), frequent, etc. collocations, finding enough common ground for a direct comparison of results is rarely possible.
This is due to not only the terms and the concepts differing in scope and theoretical commitment, but also because, secondly, the term *collocation* is linguistically massively underspecified. The only thing that seems congruent is, in Bartsch and Evert's words, "collocation as the habitual and recurrent juxtaposition of words with particular other words" (Bartsch and Evert, 2014, 48); or in the words of Paquot (2015, 460):

> "Evidence of word use in corpora has shown to an unprecedented extent that words are not isolates but rather combine with each other in preferred syntagmatic patterns to acquire meaning".

But this can of course mean anything: Collocations can be positional (co-occurring words within a span of n tokens, n-grams), in which case the term denotes contextually correlated words, but not necessarily semantically or syntactically holistic structures, like the frequent English bigram *is the*. They can be syntactically structured, as in the case of verb + noun, adjective + noun, or adverb + adjective collocations (those terms are used in Laufer and Waldman (2011); Boers et al. (2014a); Venkatapathy and Joshi (2005); Almela (2011); Wu et al. (2010); Evert and Kermes (2003); Evert (2008); Biskup (1992); McGee (2009) among others). Sometimes, even larger units like noun + of + noun are labeled collocations (Bueraheng and Laohawiriyanon, 2014; Wu et al., 2010), which would be considered lexicosyntactic or partially lexically specified constructions in more syntactically oriented approaches.
Consequentially, collocations are also rather different in terms of their syntactic and semantic characteristics and their chance of occurrence. The noun in the verb + noun collocation is an object to the noun, it forms a semantic unit with the verb in the verb phrase and cannot be omitted as regularly as the adjective in the adjective + noun collocation. Some adverbs in adverb + adjective collocations on the other hand are rather delexicalized and work merely as general intensifiers, like *ridiculously* in *ridiculously easy*, and can be considered similar to functional collocations, like two-word prepositions (*instead of*). Even within one syntactic category, collocations do not all have the same linguistic characteristics. For example, a verb + noun combination may be a light verb construction like *to give a sigh*, where the verb is delexicalized and serves as a verbifier for the noun *sigh*; a regular verb-object pair like *draw a picture*; a verb-object pair where the noun is semantically redundant to the verb like *ask a question*; it can be frequent or infrequent, like *take a break* vs. *bear a resemblance*; it can be compositional or non-compositional, like *pay the bill* vs. *jump to a conclusion*; some are very common across large parts of a language, while others are restricted to specific purposes, like *make an exception* vs. *cease*

trading; and so on (see Burger (2004) for a critical discussion with German examples, and (Wray, 2013) for a critical synopsis of unclear concepts in the various discourses).[3]

With this multitude of terms, concepts, and linguistic realizations there is so much variation in the research objects that a solid picture regarding the use of collocations as opposed to fully fixed material does not truly emerge, partially also because many studies pursue classificatory and descriptive goals.[4] Importantly, lexicography and language teaching project to the *langue*, not the *parole*: A learner is taught what is conventional in the target language, not for a target language speaker; and a lexicon that is used for translation or in a specialized classroom is expected to include a large amount of specialized and rare phrases, regardless of the percentage of native speakers that will be familiar with it. There is general agreement that native speakers use many collocations in contrast to L2, i.e. coselect words in a relatively constrained manner, but this is not usually quantified by register or native speaker cohort (see Paquot and Granger (2012) for an overview, Laufer and Waldman (2011) for an example and further references). Regardless, the unanimous agreement in the literature is summarized in this quote from Foster et al. (2014, 9 in preprint):

> "(...) language users are presumed to have a vast store of knowledge of how words in their native language (L1) most naturally combine, and this is acquired like the bulk of their L1 knowledge incidentally through a lifetime of interactions".

But some research indicates that on the contrary, coselectional constraints are a challenge even for native speakers, at least during acquisition, i.e. for children and adolescents.[5] In her detailed treatment of different kinds of formulaic language, Wray (2002), and similarly Wray and Perkins (2000), describe a u-shape in the acquisition and usage of fixed language in FLA, where children start by learning chunks and routines, and break them down into bits. Later they enter a phase of particular sensitivity to routine interactions like role play, and to longer sequences of more or less fixed language, like jokes, songs, nursery rhymes, riddles, and so on. It appears, however, that aside from these salient kinds of formulaic language, native speakers face difficulties acquiring phrasal or multi-word units like phrasal verbs or combinations of verbs and prepositions. These underly structural constraints such as (in-)separability (*look after*: *I look after her dog*, but **I look her dog after* vs. *set up*: *I set up the table*, *I set the table up*), and many have a non-compositional meaning. Crutchley (2007, 205) in a sample of 799 UK primary and middle school students finds that out of 15 phrasal verbs, 6 do not reach a 90% comprehension threshold even in 11-year-old children, and not a single one reaches a 90% threshold in the youngest age group (6;0-6;5). Even relatively frequent phrasal verbs such as *pick up*, *get over*, or *cross out* stay just below the acquisition threshold at age eleven. This is contrasted by Kerbel and Grunwell (1997, 113)'s report that

> "Contrary to the belief of six language unit teachers that they rarely used idioms in the classroom, this study reveals an average usage by these teachers of 1.73 idioms per minute";

[3] Since terms are not always clear in the literature, I will report results related to morphosyntactically and semantically diverse phenomena here (phrasal verbs, idioms, coselections).

[4] For two more in-depth treatments of different kinds of collocations, see Roth (2014) and Bartsch (2004).

[5] This is not to suggest that acquisition is constrained to younger years, but it is certainly more prominent then.

and Nippold et al. (2001), who classify up to 20% of student-directed classroom speech to contain an idiom (a semantically non-compositional chunk or collocation) in the 8th grade of a New Zealand middle school. At the same time a mean understanding of only 8.36/12 is reached in 11- and 12-year-olds. This points at a discrepancy between input, comprehension, and output: Children in the classroom are exposed to idioms very frequently, but this is not reflected in either their comprehension or their production. Idioms, phrasal verbs, and coselectional preferences are of course not all the same type of linguistic item. But if even frequent phrasal verbs are not acquired by age 11, this suggests that collocations that are held together by less than an idiosyncratic or non-compositional meaning will also be challenging for young native speakers.

In a longer curve, adolescents and young adults need to learn how to navigate a number of social routines, i.e. semantic frames, requiring specialized language. And finally, reaching higher levels of academic or professional education, native speakers need to acquire the same terms and collocations that learners are taught in the teaching of languages for academic or specific purposes. Thus, native speakers go through a highly fixed phase, then a rather productive phrase, and back to a more routinized and conventionalized use of language, certainly in specific contexts – but it is unclear that they succeed:

As one indicator that they may not, collocations are listed as challenges to the acquisition of German *Bildungssprache* and *Schuldeutsch* ('academic' or 'educated language' and 'school German') by mono- and multilingual students in German middle- and high schools in Gutzmann (2017); Beckert and Juska-Bacher (2015); Hee (2017).[6]

Secondly, there is very little research into the productive use of coselectional preferences in native speakers overall, much less in late L1 acquisition. But in one recent study Hee (2019) shows in her analysis of students in grades 5, 8, and 11 of German middle and high schools (*Gymnasium*, ages 10-11, 13-14, 16-17) that 5th and 8th graders try a number of formulations, and often stick with unidiomatic coselections like *die Flut des Nils wissen* 'to know the high tide of the Nile' (5th grade, p. 76; this cannot be reformulated idiomatically without a paraphrase like *wissen, wann die Flut kommt* 'to know, when the high tide comes in' or *die Gezeiten kennen* 'to know the tides'); Or *Höherachtung bekommen* in 8th grade, where the compound form *Höherachtung* ('higher-respect') is unidiomatic and should be *höhere Achtung* ('heightened respect') or better be verbified into *höher geachtet werden* ('to be more highly respected').

And thirdly – even for adults, inter-individual variation in L1 is not studied much overall – but in the only study into this that I am aware of, Dąbrowska (2014) reports results of a collocation recognition experiment ("Words that go together test"), where 80 adult native speakers of English of different ages and education backgrounds were prompted to select the most "natural or familiar phrase" from a list of five, for example *blatant lie* on a list with *clear lie, conspicuous lie, distinct lie, recognizable lie*. She reports a range of correct or expected responses from 28% to 98%, 11 to 39 out of 40, with a mean of 29.5 correctly recognized collocations (74%), and no significant correlations of test performance with any corpus measures like frequency or mutual information scores (MI). It should be noted that some task or prompt effect might play a role here though, because the prompt was to select the most familiar or natural, not the most restricted or otherwise collocational pair.

[6]The two notions related to register competence but also skills of linguistic thinking, abstraction, and expression are discussed in the context of discourses in German-speaking pedagogy and didactics centered around the development of teaching methods that incorporate a training of linguistic features relevant to academic or school-related registers in all subjects; and the lack of success experienced by multilingual children in German classrooms, see also Haberzettl (2016, 2009); Cantone and Haberzettl (2009).

But in one of the groups reported in the paper, *boost production* is marked as expected, but the next item on the list, *double production*, is also frequent, and in two meanings (*to double (the) production (of cars)*, *a double production (of two plays at once)*). Still, a range of 0.3-0.98 in native speakers appears enormous, especially when considering how generalized statments with respect to 'nativelike' selection are made in the literature.

This study is also interesting in another respect: High correlations between test performance and metadata are found for reading habits, grammar, and general vocabulary tests (0.51 for author recognition and 0.34 for self-reported reading; 0.43 for grammar; 0.53 for general vocabulary), emphasizing the role of linguistic input. While this could point towards a frequency effect, a number of other explanations are equally plausible: It is possible that a deeper understanding, anchoring or entrenchment of collocations promotes the enjoyment of reading; or that words are better retained if one reads more different words (more docking points – better structure – better retainment); or that a generally higher linguistic capacity allows for better understanding and recognition of conventional structures; or for both more enjoyment of literature and language in any context.

There is not much research into the function or the structure of collocations as seen from either lexicogrammar (paradigms of collocations) or from the individual word (coselectional constraint of a specific word, distributions of collocates). Of course implicit models of these exist through collocation lists and dictionaries, as well as observations of the use of collocations in learners (see section 2.2), but those are not discussed functionally or structurally.

The role of frequency can not be assessed reliably from the research as it exists to date. Although frequency effects are modeled as central to language use and learning in usage-based grammar models (see section 2.2), in the case of collocations, frequency is a bizarrely amorphous concept. Some positional collocations are extremely frequent (*is the*), some are frequent in a specific context but not most others (*usage-based grammar*), some are overall infrequent (*set ablaze*). The reason that the statistical identification and extraction of collocations has proliferated in lexicography lies in the desire to abstract from orders of magnitude and the frequencies of the individual collocates and get a unique 'overall' value for the strength of a collocation. There are mathematical problems with this, which will be discussed in chapters 4 and 7. But on a more theoretical note, if entrenchment through frequency as a fundamental structuring process is presumed, then abstracting from frequency is arguably not helpful for the model, because it negates frequency effects normalizing them into distributional effects at best.[7] It is therefore not even clear what a normalization of frequency that would be constructive to the theoretical model may look like. This is especially so because the chance of occurrence for collocates differs significantly between the different categories of collocations (for example, verbs and nouns are more correlated than adverbs and adjectives). Thus, the study of collocations holds no obvious hypotheses for a quantitative study of coselection, aside from the hint from Dąbrowska (2014)'s study that L1 variance is to be expected.

The difficulty of modeling frequency also lies in the fact that frequency as it exists in the language environment clearly unfolds its effects on the individual speakers' minds, but it is rather difficult to even somewhat coherently model the input of a specific speaker beyond toddler years, and corpora do not reflect a speaker's input either (nobody, for example,

[7] Sometimes this is countered with the introduction of the concept of salience to explain how things can be learned and retained from only being introduced or mentioned once, see Ellis (2006b, 2016). But this only extends the problem to the challenge of modeling the interaction of two highly variant and fluctuating factors without eliminating the need to also explicitly model frequency effects.

reads dozens of newspapers front to back over decades). Thus, while language on a *langue*-level appears to be coselectionally constrained, it is basically unknown at present what coselectional constraint looks like outside of fully fixed material in individual speakers or speaker cohorts, when or how it is acquired, how broadly it is acquired across contexts or registers, or how variable coselectional knowledge or constraint is in native speakers of various ages or backgrounds.

2.1.2. Collostructional distribution

Coselectional preferences on a collostructional level, i.e. in the interaction of one or more words and their syntactic environment, have been studied at the example of a number of syntactic and lexicosyntactic phenomena.

Collostructional analysis is a statistical approach that was first introduced by Stefanowitsch and Gries (2003) and extended in Gries and Stefanowitsch (2004); Stefanowitsch and Gries (2005). In their 2003 paper the authors show that verbs are associated with specific constructions, for example *give* is most strongly associated with the ditransitive; but also to certain categories of tense, mood, and aspect (TAM), where *talk* has a strong association with the progressive, or *let* with the imperative. They then move on to show that such associations exist not only between words and constructions, but also between the two slots of a construction, as in the *into*-causative (Stefanowitsch and Gries, 2005, 12):

> "In general, the results show that in the case of the into-causative, the semantic coherence between the covarying collexemes is based on conventionalized causal frame sequences, i.e. on (culture-specific) frame-semantic knowledge of what typically causes what. Take the first four pairs [*fool into thinking, mislead into thinking, mislead into believing, deceive into thinking*, AS]. All of them instantiate a relationship between a trickery frame and a belief frame. If we include all significant covarying-collexeme pairs with belief results, it turns out that this relationship is in fact the predominant one for this frame in the into-causative."

The three papers come with a whole array of case studies on constructions discussed also in CxG, showing that each construction is statistically more or less associated with some lexemes. In their interpretation this provides evidence for the association of constructions with semantic frames through lexemes, something that is modeled as entrenchment in CxG.

Similarly, Divjak and Gries (2009) show in what they call a *behavioral profile* that near-synonyms like *begin* and *start* show preferences regarding the semantic aspects of their objects (like animacy) or syntactic environments (like aspect, tense, etc.). They contrast this with a similar pair from Russian *начинаться/начаться* 'načinat'sja/načat'sja' ('to begin') showing that all four have different behavioral profiles, i.e. tend to co-occur with diverging groups of syntactic and semantic features.

With fewer claims to semantic specificity or idiosyncrasy, Römer (2005, 118–125) in her contrastive study of spoken English vs. textbook English finds that lexemes are distributed unequally over progressive tenses. *Accepting*, for example, occurs overwhelmingly in the present progressive (>90%), and never in the past perfect progressive in her corpus of spoken English, while for *betting*, occurrences in the past perfect progressive make up >14%. Wulff (2006) measures different associations for the *go-and-V* vs. the *go-V* constructions, finding an overlap of 20 types between the two groups, where go-and-V occurs with 92 more, and go-V with only 25 more. Nicolle (2009) reports differences between *go-and-V,*

come-and-V, go-V and come-V in their coselection, where come-V in particular appears much more often in imperative clauses and less in declarative clauses than the other three.

The method of collostructional analysis is not intrinsically different from measuring statistical associations between words in collocation extraction, only now applied to words prefiltered by construction:

> "Collostructional analysis always starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction (i.e. occur more frequently or less frequently than expected)" (Stefanowitsch and Gries, 2003, 214)

There are mathematical and epistemological problems with the application of statistical measures to corpus data, which will be discussed in chapter 4. But there is also another reason why those results should be viewed as exploratory rather than confirmatory: Although results from collostructional analysis and related approaches have been widely accepted as evidence of the either purely conventional or semantically framed coselection of items, they are in fact only patterns. Patterns are not epistemes (Dixon, 2012), they are only a first step in the epistemological process. Once a pattern is found and interpreted within an existing framework, hypotheses from this interpretation need to be built and tested on new data. Otherwise, epistemological uncertainty ensues, because it is never quite clear which aspect of the pattern was noise and which was signal – is the part of the pattern that is chosen for further investigation an outlier? An exemplar? A prototype? An exemplary distribution? None of the above?

All case studies listed here are performed more or less exploratively, similarly to how collocations are extracted from corpora for lexicographic purposes. Where hypotheses are made explicit, those are either very general (that there would be some kind of semantic coherence between elements, Stefanowitsch and Gries (2005)), or purely distributional (that lexemes between constructions would differ, like in Wulff (2006) or Divjak and Gries (2009)), and linguistic interpretation based on argument structure, verb type, or other categories is post-hoc.

While results may be interpretable, they are not hypothesis-based applications of usage-based theory, at least not for hypotheses beyond the intuition that lexemes are distributed unevenly across syntactic construction and co-occurrences (the *idiom principle*), or that there should be some kind of semantic coherence (which goes largely undefined). There is to my knowledge only one application of collostructional analysis in a synchronic[8] context that is hypothesis-based and specific, and that is Hampe (2011)'s study of the resultative vs. what she coins a denominative construction (of naming, branding, coining, etc. things as [ADJ]). In all of the other studies cited, post-hoc interpretations are made and results judged as plausible.[9]

[8]There is some work done diachronically (Hilpert, 2012), which may be easier to model in hypotheses because the earliest and the latest data points can be used for reference.

[9]Stefanowitsch and Gries (2005, 12) even model this as a methodological necessity:

> "What is crucial in the present context, however, is that even this relatively precise description of the construction's semantics does not allow us to predict combinations of cause and result predicates. As mentioned in the preceding section, the principle of semantic compatibility predicts that these combinations should be semantically coherent, but it does not provide us with an expectation concerning the kind of semantic coherence."

Aside from the fact that it is of doubtful plausibility that an a-priori idea (of which lexemes may be

As such, while the intuition that lexicogrammatical units form clusters across categories as formulated in Sinclair's *idiom principle* is confirmed in this line of study, it cannot be seen as very solid evidence for the existence of *specific* forces of attraction, or conventionalized grammar. Worse still, while there are many studies replicating the same general tendency, I am not aware of a single replication study of the same collostructions on another corpus that would confirm more or less *stable* forces of attraction for specific items in the first place. It is plausible that the *into*-causative is in fact associated with a TRICKERY-frame, as suggested by Gries and Stefanowitsch. But to say that it truly is, a replication on new data with this specific hypothesis is required. If in another corpus, other words and with them another frame are prevalently associated with the *into*-causative, the conceptualization as a specific force of attraction would have be reconsidered. Replication is always necessary in empirical research,[10] but it is trvially necessary if *specific* forces of attraction are implied – without replication, it is impossible to tell whether results from a single corpus correlate to any other data. Thus at present, it is unclear whether the unequal distributions of collostructional coselection are lexicogrammatical, thematic, or even random within a reasonable semantic space.

While frequency and distribution of syntactic coselections are studied much in language acquisition research (see section 2.2.3), and studies into the individual sensitivity to distributional aspects as well as priming and processing exist (see section 2.2.2), collostructional analysis and related approaches have not yet shown major interest in inter-individual or stratified variation. Little can therefore be said about what to expect from learners or native speakers about such coselectional preferences, except that if they are indeed stable, distributional knowledge requires a process of acquisition. The association of syntactic structures with specific semantics could be learned implicitly, but the study of collocations suggests that even if it was, this association may not be reflected in productive writing. Timing and ease of acquisition would likely depend on similar factors as in the case of collocations: Frequency, salience of form, semantic transparency, and functionality.

2.1.3. Idiosyncrasy

Usage-based approaches set out to explain grammar through meaning and to provide an explanation for language in use, which was understood as impossible with a syntactic and a lexical module as they were modeled in transformational grammars. Lexically unspecified syntax overgenerates where it comes to idiomatic restrictions, and over- or undergenerates where lexically partially specified constructions are used productively with novel material (like the LET ALONE-construction, Fillmore et al. (1988)). As constraining forces on the productive potential of such constructions, Goldberg (1995, 2005) stipulates the *Semantic Coherence* and the *Correspondence* principles. Those state that there must be semantic compatibility between the argument role of a construction and the participant role of the verb; and that the semantically salient roles of an event require encoding in a verb-argument realization. In the same spirit, Levin (1993, 1) expresses the idea of a semantic mapping for verb-argument structures and verb senses:

> "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large degree determined by its meaning".

semantically coherent with a syntactic or semantic frame) is impossible to formulate, this is also prone to confirmation bias and result bending. Even if one were to accept the prediction of collexemes as an impossible endeavor for a first study, hypothesis-based replication should still be possible.

[10] See Plonsky (2014) for a synthesis of arguments around research quality and replication in linguistics.

The idea that verb senses correlate with verb-argument structures is an integral part also of valency grammar and frame semantics (Herbst, 2014a,b; Engelberg et al., 2015; Engelberg, 2014; Martin, 2008; Boas, 2011; Boas and Dux, 2017). More recently, however, growing evidence from those approaches shows that neither lexical nor frame or argument structure semantics alone suffice to explain the variation of verb + verb-argument structure coselection as it is found in corpora. Verb lexemes often fail to appear in verb-argument structures that they should be licensed to occur in semantically; or occur in verb-argument structures that they should not occur in; and the same verb-argument structures can take different meaning depending on the verb lexeme: *They built the house on a bad foundation* vs. *They carved a toy on a couch* (Boas, 2011, 213). It appears that these are not isolated cases, but a relatively widespread phenomenon even for groups of semantically very similar verbs. At the same time it is not entirely clear what constitutes similarity between verbs either.
Boas (2011) analyzes syntactic alternations of 35 *build*-verbs that are defined as one semantic class through their syntactic behavior in Levin (1993). He shows that a syntactic classification leads to an exclusion of verbs that by any standard of semantic similarity would be included (like *construct*, because it cannot occur in the resultative alternation **Lena constructed the bricks into a building*), while some rather strange candidates like *grind* are included. Boas goes on to suggest more fine-grained semantic classes defined by more specified frames. As such, what appears as an idiosyncrasy may sometimes be a result of coarse categories. In fact, it may be that a very specific verb is coselectionally constrained simply because its meaning evokes frame participants that are also limited in the world we live in. For example, the verb *meow* will typically only allow for meowing subjects. As Plank (1984) argues, there are whole classes of verbs with a very specific meaning, like the rather restricted German verb *knacken* in its transitive sense, 'to crack', whose objects are limited to nuts and, in a meaning of 'to solve', difficult riddles or cases.[11] In this case, extreme coselectional constraint looks like an idiosyncratic pattern but is not actually determined by preferences of the item choosing only one out of many possible collocates, but by a restriction of the potential set of collocates due to its high specialization.

However, it appears that even for near-synonyms and more general verbs, idiosyncrasies and failure to occur in some verb-argument structures over others exists. Faulhaber (2011) compares 87 verbs by how much of an overlap semantically similar verbs have regarding the syntactic structures they appear in. For example, out of the group of *answer, reply* and *respond*, *answer* can appear in a ditransitive construction with an indirect object (*to answer someone's question*), but the other two verbs require a prepositional object (*to reply to someone's e-mail).* Faulhaber estimates that about 80% of the verbs in each of her groups behave in accordance with semantic aspects, while some 20% show idiosyncrasies that cannot be explained without accepting conventionality as a cohesive force. Similarly, Dux (2016) analyzes argument structure distributions of semantically narrowly related verbs of theft (*steal, snatch, pilfer, embezzle, shoplift*) and change (*alter, change, modify, transform, turn*) in a frame-semantic analysis. He concludes that

> "(...) verbs with nearly identical meanings exhibit significant variation in their distribution across syntactic contexts (i.e. valency constructions). For instance, among the five English Change verbs classified together in both Levin (1993) and FrameNet (Fillmore and Baker 2010), no two verbs occurred in the

[11] Perhaps metaphorically also 'to get through the shell of unapproachable people'

> same range of valency constructions with similar frequencies and, more strikingly, no valency construction was found to occur with all verbs of the class in the data set" (Dux, 2016, 427).

In summary, it appears that while verb-argument structures seem to correlate with semantic meaning to a degree, they do not paradigmatically alternate with their respective frames for all lexemes. The question then is what constraints such syntactic alternations underly.

With a different focus, but a similar observation of idiosyncratic tendencies, Zeldes (2013b, 2012) finds that the potential productivity of derivational morphemes in German and verb-argument coselection in English cannot be attributed to only to semantic classes, formal semantic meaning, or world knowledge. Rather, he suggests that "[p]roductivity is at least partially an idiosyncratic, language-specific property" (Zeldes, 2013b, 137). This is relevant for the study at hand because productivity is in a sense the negative of coselectional constraint, since an item is more productive if it allows for more, and new, combinations.[12] If the coselectional behavior of verbs and verb-argument structures cannot be predicted from semantic or syntactic rules, then it means they must be acquired at least partially in an item-based fashion or distributionally with clusters of items. Item-based learning is described in detail for early FLA, where one of the interesting aspects is the necessary re-organization and re-structuring of elements where previous generalizations have to be limited. What is particularly intriguing here is that both the verbs and the verb-argument structures involved are not idiosyncratic per se, but their combination or combinatorial power appears to be. This means that, if the process is truly marked by arbitrary idiosyncrasies and not guided by external forces (such as semiotics, phonotactics, or mnemonic salience), an interplay of very detailed implicit hypotheses and rather strong general rules as they are taught in the classroom is to be expected. Some of this will be discussed at the beginning of chapter 3.

Again, little is known about such idiosyncrasies in productive writing of an individual or smallish cohorts, or the awareness and receptive sensitivity to them in native speakers of different strata. Thus, hypotheses can only be derived against the idealized corpus speaker as in the previous sections.

2.1.4. Summary

Empirical studies in usage-based approaches mainly from collostructional analysis and frame semantics have shown a tendency of lexemes towards an unequal distribution across argument structures, constructions, and morphosyntactic categories in general. It is robustly found across studies that coselections are not randomly distributed by virtue of the frequencies of the involved items, i.e. that language as it is represented in corpora is not as freely combined as the *open choice principle* might suggest. However, while the data usually allows for conclusions of either semantics or arbitrary idiosyncrasy as the guiding forces of this, hypothesis-based replication is necessary to confirm that these reflect stable and specific forces of attraction between items, i.e. that these observations are indeed evidence of the *idiom principle*. It is currently unknown how specific these forces of attraction are in the *langue*, and whether they are driven by higher-order forces (like frequency, salience, semantics, semiotics, or other aspects) or truly idiosyncratic (arbitrarily attached

[12] The two are not entirely complementary though, because native speakers are known to be both more productive and more coselectionally constrained. An exact model of productivity vs. coselectional constraints does not exist to date.

to individual lexemes). In any case, they involve abstract linguistic categories and are thus clearly functionally different from chunks (for example, they are not ad-hoc communicative bits, but require syntactic and contextual embedding). If coselectional constraints are indeed stable and specific, this would mark a distributional phenomenon (each item comes with a distribution of preferred coselections) rather than simple form-meaning mappings as in fixed language, and distributions may also differ by registers or word senses, making coselectional constraint a complex relational phenomenon at the center of lexicogrammar.

While convention and idiosyncrasy have been identified as an aspect of coselections across a number of lexicosyntactic levels and phenomena, the extent of this in the *parole* of native speakers is essentially unknown, as is the timing and process of acquisition, and the variability among native speakers or contexts. First studies at the example of collocations indicate challenges in the productive use and acquisition of such idiosyncrasies in L1 at least until early adolescence and a high variance in adult native speakers. This may be due to the sheer number and combinatorial power of potential item-based constraints, a less fixed and salient form, and higher communicative specificity (i.e. lower applicability or general functionality) of coselectional constraints vs. chunks. Thus, findings from corpus studies of collocations, collostructions, and idiosyncrasies in verb-argument structures and coselections suggest that a theoretical account of coselectional constraint in its own right is due.

2.2. Coselection in language learning

In acquiring language, children first learn communicative chunks that are then broken down, rearranged, and generalized to a functional grammar. Detailed accounts of children's generalization from chunks and lexicosyntactic islands have been provided by Kuiper et al. (2009); Tomasello (2000, 2009); MacWhinney (2014, 2004); Lieven et al. (1997); Bates and Goodman (1999), and others. These observations were made in sharp contrast to the *poverty of the stimulus argument* stipulated by the generative paradigm (see Cook (1991) for a synopsis), and the usage-based framework developed in what could be described as a strong resistance to the generative paradigm as a whole (Tomasello, 1995; Harris, 1995). This is why a central emphasis has been placed on the learnability of natural language from general cognitive capacities, and without language-specific a-priori knowledge. Specifically, the processes of chunking and segmentation on a phonetic, phonological, and lexical level, as well as the processes of abstraction and generalization in morphosyntax, and a sensitivity to frequency and distribution are suggested as main driving forces of language construction (Ellis, 1996, 2008; Ellis and Simpson-Vlach, 2009; Ellis et al., 2014; Plunkett and Marchman, 1991; Tomasello, 2000; Goldberg, 1995, 2006; Alishashi and Stevenson, 2005; Bowerman, 1982; Bates and Goodman, 1999; Braine, 1987; Naigles, 2002; MacWhinney, 2004; Perek and Goldberg, 2017, and many others).

Although there is some debate on whether those processes are equally active in SLA, this focus is most manifest in the model of a phraseological continuum that is widely accepted in both FLA and SLA research and in usage-based grammar models:

> "(...) there is general agreement that phraseology constitutes a continuum along which word combinations are situated, with the most opaque and fixed ones at one end and the most transparent and variable ones at the other (...)." (Granger, 2005, 1, see also for references to more similar continuum models);

More will be said about the continuum hypothesis in section 2.3.1. For now, it is introduced

as a reference point for the discussion of coselectional constraint in the context of language learning.

> "It is essential to see categories as forming a continuum from the most free combinations to the most fixed idioms, rather than discrete classes. Dividing lines cannot be strictly drawn, though points along the scale are regarded as somehow reflecting psychological reality (...)" (Howarth, 1998, 35).

With the acceptance of the absence of categorial boundaries, coselectional preferences are not typically modeled separately in language learning research, but simply placed somewhere in the middle of the continuum. However, this appears to be misleading in at least three ways:

Firstly, if the continuum is conquered from the fixed end to the freely combined, learners in L1 and L2 should first acquire chunks, then slowly dissolve them into collocations, and only then gain productive skills. But this is not what happens – rather, both L1 and L2 learners move from the fixed end of the continuum to the freely combined, and only then begin to specialize and conventionalize. This is also what would be expected from a functional perspective, which is emphasized in much of the FLA research,[13] but not applied consequently in the continuum hypothesis: In a hierarchy of linguistic needs, basic communication skills come first, then the need to extend communicability to many situations, while conventional and eloquent communication and precise self-expression would be ranked lower in priority. A reflection of this is found in the study of collocations in SLA, where a general consensus is that some frequent and general native-like collocations are overused, while rarer and more specialized collocations are underused, and that the details of coselectional constraint are usually not acquired even by advanced learners. Some of this research is reviewed in section 2.1.1.

Secondly, coselectional preferences have much more complex distributional aspects than chunks, like groups of preferred coselections out of which some are more prototypical, frequency distributions of both coselected items, and triple coselections like coselections of a syntactic and two lexical elements, rather than a simple form-meaning mapping. This means that much higher complexity and opaqueness of precise acquisition comes at a much lower immediate communicative benefit, which should have some repercussions on the acquisition model.

And thirdly, coselections are not always, and some are not usually, contiguous, mnemonically friendly chunks, and it is unclear in how far processes of segmentation and generalization apply to them in the same way as to chunks. These aspects will be discussed in sections 2.2.2 and 2.2.3.

2.2.1. Learner corpora, collocational competence and phraseological complexity

In learner corpus research, similarly to what was reported in the earlier section on collocational study, there is much research of different types of collocations and little theoretical integration of the different strands.[14] However, the general consensus is that colloca-

[13] With high emphasis by Tomasello (1995, 2009); Ellis (2006a), but also in all of functional linguistics and where language is modeled as a complex adaptive system, e.g. Five Graces Group et al. (2009); Tucker and Fawcett (1996); Ellis (2016).

[14] "Any survey of the literature highlights what may, at first sight, look like contradictions, but is in reality nothing but a consequence of the considerable heterogeneity in both data and methods" (Paquot and Granger, 2012, 11 in preprint).

tions are somehow difficult for learners, and perhaps more difficult than other aspects of acquisition:

> "The overall picture that emerges on the basis of collocation studies is that the use of collocations is problematic for L2 learners, regardless of the number of years of instruction they have received in L2, their native language, or type of task they are asked to perform" (Laufer and Waldman, 2011, 651);

This difficulty is studied mostly in over-/underuse studies and in error analyses of learner text compared to L1 text, where

> "the overall picture that emerges from learner-corpus-based studies is that learners' use of co-occurring combinations is characterized by a mixture of underuse, overuse and misuse" (Paquot and Granger, 2012, 12 in preprint).

The mixture is not in fact a mixture though, but a categorial separation by different kinds of collocations. (Durrant and Schmitt, 2009, 158) summarize that

> "[t]he general picture that has emerged (...) is that advanced learners do appear to use formulaic language (...), but often not to the same extent as natives (...). At the same time, learners tend to overuse (in comparison to native norms) a small range of favourite phrases, especially if they are frequent/neutral items or are a cognate to L1 forms (...)",

The authors go on to show that learners even at advanced stages underuse collocations compared to native speakers, and particularly those where both collocates have low corpus frequencies, i.e. collocations with high statistical association as measured by mutual information scores; which is replicated in essence in Granger and Bestgen (2014) and even suggested as a measure for language assessment in a dimension labeled *phraseological complexity* (Paquot, 2019; Paquot et al., to appear; Paquot, 2018). The observation of an overuse of frequent items has similarly been made in the concept of a *lexical teddy bear* (Hasselgren, 1994), a word or a phrase that learners learn early and then hold on and use disproportionately often even much later in their acquisition process. The term was later extended to a collocational (Nesselhauf, 2005, 69) and phrasal teddy bear (Ellis, 2012b). Regarding the error-proneness of collocations, Laufer and Waldman (2011, 647) in their study of Hebrew-L1 learners of English note that

> "learners at all three proficiency levels produced far fewer collocations than native speakers, that the number of collocations increased only at the advanced level, and that errors, particularly interlingual ones, continued to persist even at advanced levels of proficiency".

Similarly, Nesselhauf (2003) in her study of the use of verb + object collocations in German-L1 learners of English finds that unconventional coselections (non-occurring in BNC or not accepted by native speakers due to ungrammatical prepositions or semantic incoherence) make up about a quarter of all noun + object uses in those essays. Importantly, Nesselhauf (2005) shows that error rates barely change when learners are allowed to use dictionaries, suggesting as Laufer and Waldman (2011) discuss that learners are not experiencing problems with retrieval during writing, but are unaware of the problem.
Coselecting words that are not idiomatically coselected in the target language is only one kind of error that can be made in this context, with the particular case of the exchange of

one collocate for a near-synonym that does not collocate well with the other word; others include semantic overextension with a particular sensitivity to L1-transfer, and grammatical errors in fixed chunks (see Laufer and Waldman (2011, 652ff.) for a discussion and references). Those are interesting because they all point at the different character of coselectional constraint or coselectional preferences vs. chunks. In fact, both the observation of lexical, collocational, etc. teddy bears and the fact of a general overuse of frequent collocations in learners suggests that the more fixed a collocation, the easier it is to learn, retain, and use for learners. The same is suggested by Nesselhauf (2003, 233):

> "The lowest rate of mistakes, on the other hand, is found with combinations classifed as RC1 [very restricted, AS] (such as *pay attention* or *run a risk*). It therefore seems that whereas learners are mostly aware of the restriction in combinations where the verb only takes a few nouns, they are less aware of restrictions in combinations where the verb takes a wider range of nouns (such as *exert, perform*, or *reach*)".

Form, when it is fixed or salient, generally seems to facilitate the learning of collocations, as has also been shown in the study of alliterative collocations (*slippery slope*). Those are generally better retained when the learners' attention is directed to this feature and slightly better in short-term memory even without awareness-raising (Boers et al., 2012, 2014b).

Another aspect of form is the syntactic realization of a coselection. There, it appears that aspects of more general syntactic development seem to overwrite native-like coselections of syntax and lexis, as is suggested by Güngör and Uysal (2016) and Pan et al. (2016), both being corpus analyses of professional writing (academic texts and professional telecommunication articles written by Turkish-L1 and Mandarin-L1 authors respectively, both compared with native speaker texts from the same register). Both studies find an underuse of NP- and PP-based lexical bundles (*the nature of, for each of the*) in favor of clause- or VP-based coselections (*it was found that*, examples from Güngör and Uysal (2016)) in learners. This can also be interpreted as a lack of structural register competence or a failure to handle higher information density, since nominal style is representative of conceptually written and, specifically, academic registers:

> "(...) the distinctive communicative characteristics of academic writing (informational prose) have led to the development of a discourse style that relies heavily on nominal structures, with extensive phrasal modification and a relative absence of verbs" (Biber and Gray, 2011).

These studies are among the first to provide evidence to the existence of diverging coselectional preferences on not only a lexical, but also a structural level in learners vs. native speakers. This means they are much more complexly interwoven with other aspects of language and do not seem well described as nearly identical to chunks, only somewhat less fixed. These results also show that coselectional preferences in learners can likely not be modeled independently on a separate dimension, but require an interpretation in the context of chance of occurrence and developmental maturity of all categories involved. For example, learners tend to generally underuse modifiers (Hirschmann et al., 2013; Hirschmann, 2015; Vyatkina et al., 2015), which means that the number of modifying collocates is also more restricted; or, in a dialectical interpretation, with a higher number of collocations, the gap in modifiers may close. The same applies to coselections of a structural kind. A verb can only be coselected with the ditransitive when the structure

is sufficiently acquired, unless the distransitive is not analyzed but learned as a chunk.[15]. Another aspect that requires attention is a functional chance of occurrence or the necessity to use coselectionally constrained items. Since little is known about the inter-individual or intra-individual variance of collocation use or coselectional constraint in L1, it is not uniquely possible to define a standard which learners should aspire to adhere to. It is possible that learners in using fewer collocations write texts that also differ in other respects, and that those texts do reflect an L1 standard, just a different one. A similar argument is brought forth in Lambert and Kormos (2014) for measures of complexity, accuracy and fluency (CAF) in language assessment.[16] This then suggests that, eventually, a variationist analysis is needed to fully capture the effects of coselectional constraint in language learning, one that considers register-specifics, developmental constraints of learners, and L1 variance as well.

2.2.2. Distributional sensitivity

Coselectional constraint is a distributional phenomenon, and if it is to be learned implicitly, a distributional sensitivity on the learners' side must exist. The concept of a distributional sensitivity concept that relates to that of categories: Items of a category are distributed in certain ways, and each occurrence of an item may influence the perception of the category as a whole. This is a general sensory process that can be harvested for learning.

> "When a listener hears many good examples of a /b/ in a row, they are less likely to classify other sounds on, e.g., a /b/-to-/d/ continuum as /b/. This phenomenon is known as *selective adaptation* and is a well-studied property of speech perception" (Kleinschmidt and Jaeger, 2016, 678).

This process can be seen as a progressive prototypization,[17] meaning that if a hearer hears many good /b/ sounds, the prototype of a /b/ is sharpened and differences to other sounds stand out more clearly, in the same way that the sensory perceptions are sharper in a sensory deprivation environment; how hues of colors are differentiated more clearly in monochromatic paintings; or how microtonal pitch differences can be distinguished in pieces of music that range less than one whole tone.[18] If a prototypical exemplar is

[15] This observation was made for FLA by Diessel and Tomasello (2001) with respect to early learned verbs that occur with clause type objects, like *see* or *think*, where subordination could not be extended to new contexts in young children yet. Consequentially, they would be unable to coselect the syntactic structure with other verbs, because the structure is not yet available for selection at all. Rather, *see* or *think* and the subordination are selected as a chunk.

[16] "Due to the complex and dynamic nature of the variables involved in the developmental process, however, local fluctuations in accuracy, fluency, and syntactic complexity will not provide adequate insight into task-based SLA. Without theoretical modelling and empirical support linking performance measures to the use of developmentally more advanced language, task-based research is likely to result in mixed findings that are of limited value for SLA", (Lambert and Kormos, 2014, 6).

[17] For an overview of prototype and cognitive category theory, see Rosch (1983); Lakoff (1987); Geeraerts (1989); Lakoff (1999).

[18] Another example of the same process, but both productively and outside of the auditory or acoustic sphere, is the following excerpt from an portrait of basketball player Dirk Nowitzki in *Zeit Magazin* No. 46/2019, a supplement to the weekly newspaper *Die Zeit*:

> "Over time, I have seen so many workout sessions of him an Geschwindner [his coach, AS], so many throws, that I basically know the order by heart. It is always the same, it is always different. It took a while for me to realize what the two of them are actually doing. That it is not about the exercises themsleves, but about the tiniest details of their execution. At the tenth time the monotony has a meditative effect, at the 25th time one

entrenched through repeated exposure, the category is acquired as a relational triangle of category boundaries, prototypical exemplars, and atypical exemplars. This means that distributional learning is a process of continuous comparison of the exemplars of a category, and thus a relational learning process.

Distributional sensitivity is not a uniquely human characteristic, as it has been shown to exist in rats (Pons, 2006) and songbirds (Comins and Gentner, 2015; Fehér et al., 2017). Both species are interesting cases because rats are highly intelligent and social animals, while songbirds develop groupwise acoustic patterns and even, to a degree, song culture; both of which are aspects that can be mapped to the linguistic development in humans. However, it also shows that distributional learning is not language-specific, but a general learning mechanism that is related to *Gestalt*-principles of perception, which are equally relational (through part-whole relationships, foreground-background perception, and relationships to previous perceptions). As Aslin and Newport (2014) point out, distributional learning has been shown to provide good models of syllable and word segmentation and statistical patterns of syntax; and a number of computational models have shown that syntactic patterns, verb-argument structures, and semantic classes of verbs can all be learned from distributional cues alone (Aslin and Newport, 2014; Alishahi and Stevenson, 2008; Redington et al., 1998; Resnik, 1996).

Furthermore, distributional sensitivity in learning novel syntactic constructions has been shown among others by Casenhiser and Goldberg (2005), who show that children and adults acquire constructions faster if those are presented to 80% with one new verb and the other 20% distributed on many other new verbs, which approximates a Zipf-distribution as it naturally occurs in language and in child-directed speech specifically (Ellis et al. (2014, 61), Goldberg et al. (2004), Zipf (1965, 47)).

Importantly, the concepts of frequency and distribution should not be conflated. A distribution is made from frequencies, but this may not necessarily be what is noticed about it. Rather, a long-tailed distribution with several highly frequent and any number of infrequent items provides a perceptual scaffolding for category learning through the repeated priming of prototypical exemplars, where it may not matter much if a prototype is 10 or 100 times as frequent as a marginal exemplar. Thus, a central member of a category can be identified either by being the most frequently occurring, or by pre-defined features. In other words, if the category is already known, a prototypical exemplar can be identified through this knowledge, but if it is not yet well understood, the features of a frequently occurring exemplar can be used to mark out both the central features and the boundaries of the category. This is also implied in the concept of preemption or negative entrenchment, i.e. a blocked association through the total absence of evidence (Boyd and Goldberg, 2011; Stefanowitsch, 2008), which is important for learning because it constrains

begins to grasp the ever-same as a mosaic of the tiniest variations in details. On a rough level, everything stays the same, so that one may monitor what works on a fine-grained level" (my translation);

The German original reads:

"Über die Jahre habe ich so viele Trainingseinheiten von ihm und Geschwindner gesehen, so viele Würfe, dass ich die Reihenfolge im Grunde auswendig kann. Es ist immer gleich, es ist immer anders. Es hat eine Weile gedauert, bis ich dahintergekommen bin, was die beiden da eigentlich tun. Dass es nicht um die Übungen an sich geht, sondern um die winzigsten Details bei ihrer Ausführung. Beim zehnten Mal entwickelt die Monotonie einen meditativen Effekt, beim 25. Mal beginnt man, das immer Gleiche als Mosaik winziger Detailvarinten zu begreifen. Im Groben ist alles gleich, damit man überprüfen kann, was im Feinen funktioniert".

possible overgeneralizations.[19]

Learners are also particularly sensitive to a certain level of specificity which is called a *basic level category*. For example, the words *dog* or *wolf* are labels of basic level categories, while the words *Dalmatian* or *manewolf* are not, and neither are *canine* or *vertebrate*. Basic level categories are the prototypes of categories, and they are particularly salient and persistent in learning (Eimas and Quinn, 1994; Mervis and Crisafi, 1982; Markman and Wisniewski, 1997; Tanaka and Taylor, 1991) and facilitate the learning of new words (Klibanoff and Waxman, 2000; Callanan, 1989; Emberson et al., 2019).

Exemplars, prototypes, and basic level categories or labels are not all the same thing, but they all point to the relational character of learning. Item-based learning is never truly item-based as in limited to information about a single item, but appears to be correlated with many other items instantaneously.

How does this translate to coselectional preferences though? It has been mentioned previously that those do not appear to be as easily learned, certainly not by children as young as those who acquire the syntactic patterns simulated in computational models, and even a formal description appears difficult in present day linguistics (see section 2.3). There is not much research into the distributional acquisition of coselections, aside from the prevalent observation that collocations are not easily picked up from input unless they are salient in form (alliterative, for example). Regarding the coselection of constructions and prototypical verbs, however, there is evidence for a distributional sensitivity to coselectional constraints in native speakers and learners alike. Ellis et al. (2014, 91) in a gap fill task experiment with 285 native speakers of English found that:

> "(...) when fluent language users generate the verbs they associate with the V slot in sparse VAC frames such as 'he _____ across the ...',[20] 'it _____ of the ...', etc., their responses are determined by three factors:
> 1. Entrenchment – verb token frequencies in those VACs in usage experience; 2. Contingency – how faithful verbs are to particular VACs in usage experience; 3. Semantic prototypicality – the centrality of the verb meaning in the semantic network of the VAC in usage experience".

Similar claims are made in Ellis (2012a); Ellis and Ferreira-Junior (2009); Bybee and Hopper (2001); Ambridge et al. (2008), and similar results shown in learners by Römer et al. (2014), from experimental evidence and from written and spoken corpus data.[21] High-intermediate learners in this study were found to possess selectional preferences for verb-argument constructions which were centered around a prototypical and frequent verb, and that verb distributions were statistically similar to L1 as reported in Ellis et al. (2014).

[19] Constraining overgeneralization has been discussed as a basic necessity for category learning, because it limits the search space of meaning in novel objects. See also the discussion of nonce words like *wug* in Quine's paradox (how do children know what part of the novel object they should assign the new word to?). See Warren (2017) for a more recent synopsis. The article is fittingly entitled "Truth by convention".

[20] *He _____ across the ...* is described as construction here, because the verb slot could be filled with a non-word while the meaning of movement across a surface or space would still be deriveable from the syntax alone: *He mandools across the ground*, (their example, Ellis et al. (2014, 56)); the same could be said of the second slot: *He mandools across the tiki* (my example).

[21] Written: ICLE, International Corpus of Learner English, Granger et al. (2009), argumentative essays; Spoken: LINDSEI, Louvain International Database of Spoken English Interlanguage, De Cock et al. (2009), informal interviews on an everyday topic such as a movie the learner has seen or a country they have visited.

In two more studies, Gries and Wulff (2005) look into distributional sensitivity in the coselection of lexemes and verb-argument constructions, while Gries and Wulff (2009) test for distributional sensitivity in verb + infinitive vs. verb + gerund complements (for example *I remembered to fill out the form* vs. *I remembered filling out the form*, their example, p. 165). The latter is particularly interesting for the question of coselectional preference, because completing a verb-argument structure with a fitting verb draws not only from linguistic, but also from real-world distributions. If learners, for example, fill the verb slot of the verb-argument construction *v with n* with the most prototypical verbs *work* or *deal*, this reflects not only a linguistic choice, but also the entrenched frequency of dealing with things or working. All of these studies suggest that prototype and constructional knowledge exists in learners, which is generally explained through distributional learning and frequency effects:

> "(...) usage-based models assume that speakers retain memory traces of how verbs and other words have been heard used, generalizing these memory traces so that information about the frequencies of particular usage patterns constitute part of our knowledge of language" (Robenalt and Goldberg, 2016, 62).

However, the situation is in fact less clear than it may seem for coselectional preferences and most of all for collocations. Firstly, the examples used in productive studies like the ones referred to in this section elicit or teach prototypes in structures that have low combinatorial power. For example, one structure is presented, and while any verb can be filled in, only the most prototypical ones are actually used by participants, reflecting also the lack of communicative expression of the situation. Thus, the combinatorial power that exists in the language as it exists in the world never quite unfolds in experimental settings. The same is true of any coselection with one of the regularly studied syntactic structures – there just are not many of them, so that in analysis like Dux (2016), five words in a group can be distributed over perhaps one or two dozens of argument structures. But lexical coselection is infinitely more combinatorially powerful, and this is not an overstatement if one considers productivity. This holds a problem for entrenchment and for distributional learning: Even in larger corpora, the median frequency is 1 or 2 (*hapax* and *dis legomena*). Since more than half of the words occur only once or twice, they cannot have a coselectional set that is Zipf-distributed, because they can only co-occur with one or two words of a respective category. This also means that their entrenchment through frequency is doubtful, and a reliance on Zipf-distribution alone would not help.[22] Of course not every speaker knows every word in a corpus, and text corpora are in fact more lexicographically representative of a language than related to individual mental lexicons. Yet this only means that some words that occur somewhat frequently in corpora are likely to occur even less frequently in a speaker's environment, leaving an even larger explanatory gap.

[22] It is by the way not clear that words are Zipf-distributed at all: Williams et al. (2015) shows that while phrases (n-grams) are Zipf-distributed, words are not across orders of magnitude; And Piantadosi (2014) discusses that while Zipf-distributions occur in a number of unlinguistic areas, like music or programming languages, but also in a number of physical and biological systems, they do not actually capture word or meaning distributions well. In fact, it has even been suggested by Aitchison et al. (2016, 1) that the Zipf-distribution as such is merely an artifact of complex dynamics:

> "Recently, methods from statistical physics were used to show that a fairly broad class of models does provide a general explanation of Zipf's law. This explanation rests on the observation that real world data is often generated from underlying causes, known as latent variables. Those latent variables mix together multiple models that do not obey Zipf's law, giving a model that does".

A similar argument is made by Schmid (2010) in a criticism of entrenchment through frequency as a main driving force of associative language learning. While it is known that with formal or functional salience, collocations or idioms can be learned at first sight, if frequency was the driving force for entrenchment and entrenchment was word-wise, an unrestricted collocation of two rarer, but not extremely infrequent words would be very hard to learn: With each independent occurrence of one of the collocates, entrenchment would be weakened for the target collocation. Similarly, while lexically partially specified constructions are studied much in the usage-based literature, TRY-AND-V, LET ALONE, or the INTO-CAUSATIVE are of course much less frequent compared to a prepositional phrase or SVO word order.

Again, what is rarely seen, cannot be distributionally learned, or else confusion ensues. But generally, neither children nor adults seem to have much trouble accepting evidence from a single exemplar and making distributional assumptions from it. This has recently been shown by Emberson et al. (2019) in an experimental setting with 4- and 5-year-olds and adults, who were shown typical vs. atypical exemplars of categories like birds or fish, and nonce words to label them, and then asked to select as many exemplars of the same category from a list. The list contained other fish and birds respectively, but also other animals, flowers, etc. Children and adults extended the label readily to other exemplars of the category if they were shown a typical exemplar, but not when they were shown the atypical exemplar. They did not extend the label of the typical exemplar if they were shown several similar exemplars (three similar images of a sparrow, for example). This suggests that frequency helps to *narrow down* a category, but it is not necessary for a mapping in the first place.

In fact, it may even be that frequency by itself is not of particular help where other cues like salience, functionality, or distribution are not readily available: Nguyen and Webb (2017) in an experiment test Vietnamese-L1 English majors for collocation recognition and comprehension, where all participants have had over seven years of language training at high school and university levels. Collocations were of words in the 1000, 2000, and 3000 frequency ranges and also belonged to the same frequency ranges as collocations. Participants on average score less than 50% on all levels, and the recognition rate depends more on frequency of individual collocates than the collocation.

The idea that frequency, distribution, and conventionalization should be modeled separately is also picked up by Schmid (2015) in his *Entrenchment and Conventionalization Model*, where he defines entrenchment "as the continuous routinization and re-organization of associations, depending on exposure to and frequency of identical or similar processing events, subject to the exigencies of the social environment"; and conventionalization "as the continuous mutual coordination and matching of communicative knowledge and practices, subject to the exigencies of the entrenchment processes taking place in individual minds". This is an extension of the concept of a *conceptual frequency* that is a function of salience and functionality, rather than numerical occurrence, discussed in Schmid (2010). Corpus frequency as a main driving force in language acquisition and use has also been questioned in other works, see for example Krenn et al. (2001); Jolsvai et al. (2013); Hashimoto and Egbert (2019); Gollan et al. (2008).

Thus, Bybee (2002, 112)'s "items that are used together fuse together", while certainly descriptive of important processes in language learning and use in general, might not provide a comprehensive explanation of the learning of coselectional constraints in two ways: Items that are used together do not necessarily fuse, at least in learners; and items that are fused together may not have been used together a whole lot, as in the case of

native speaker coselections of rare items. Also, frequent items like light verbs fuse with some coselections (*make an exception, a list, a donation*) but not all (*make a playlist, a teddy bear* which seem more freely combined than the previous examples); and despite their general semantics are still selective about their collocates (**make an experience*).

A particularly interesting study in this context is Wonnacott et al. (2017), where the authors performed an experiment with children and adults teaching them artifical language islands, i.e. particle placement with specific nouns. While both groups showed better learning in a skewed input over an equally distributed input, adults were also sensitive to the specific nouns that were introduced with diverging particle placement, while children were not. This suggests that for adult learners, lexical idiosyncrasy is salient and expected, while for children it is not. A similar observation was made in an acceptability study of dative overgeneralization errors (**she said her "no", *they carried them the books)* by Ambridge et al. (2012, 2014) where adults and children of ages 9-10 showed effects for narrow-range semantic properties of the verbs, while younger children did not. Relevantly to the discussion of coselectional constraints, not verb frequency, but rather morphophonological aspects predicted ratings essentially without fail: "For any given verb, the degree of preference for a PO- over a DO-usage (or vice-versa) can be predicted almost perfectly by its semantic and morphophonological properties". This calls into question to what degree coselectional preferences are indeed arbitrary at all. Possibly they are not, but simply follow more fine-grained semantic cues than are generally modeled, and phonotactic regularities of the language.

A lack of sensitivity to seemingly arbitrary or idiosyncratic coselectional preferences in younger children may be of developmental advantage for achieving full productivity.[23] This is similar to the overgeneralization of morphosyntactic categories, like past tense morphology, but on a much wider and structural scale. However, since little is known about a) the stability of distributions (for two skeptical accounts see Schmid (2010) and Piantadosi (2014)), and b) the degree of knowledge, understanding, or usage of coselectional constraints in individual speakers, a mapping of distributional properties to distributional sensitivity in coselectional constraints cannot be derived from the present state of the research.

2.2.3. Coselection and interlanguage

What can be said about coselectional constraint in the context of a systematic account of SLA? Generally, while there are many empirical studies, and usage-based linguistics aspires to create a single account explaining both FLA and SLA, there is not much theorization in SLA research at present. Often, a *target deviation perspective* is taken, as Klein (1998, 535) critically notes:

> "[A] learner's performance in production or comprehension is studied not so much in its own right, as a manifestation of the learner's capacity, but in relation to a set norm; not in terms of what learners do but in terms of what they fail to do. SLA research considers the learner's utterances at some time during the process of SLA to be more or less successful attempts to reproduce the structural properties of target-language utterances. The learner tries to do what the native speaker does, but does it less well".

[23] See also Kempe et al. (2015, 247): "(...) children's processing limitations affecting working memory capacity and executive control constrain the ability to represent and generate complexity, which, in turn, facilitates emergence of structure"

This perspective arose from necessity in teaching as much as a generative bias in theoretical linguistics that persisted for decades. In the generative paradigm, second language learners were not thought to acquire a language (like children do in FLA) as much as to learn or study it in a more conscious and rule-based process, since access to universal grammar (UG) was analyzed to be partially or fully blocked after adolescence (O'Grady, 1996; Cook, 1985; Borer, 1996; Hilles, 1991, among others). But it is not only the generative paradigm that is to blame for a lack of independent theorization. In fact, Selinker (1972, 209f.) early on provided a framework in which learner language could be modeled in its own right and explicitly without reference to a generative or UG assumption:

> "It is also important to distinguish between a teaching perspective and a learning one. (...) This paper is written from the learning perspective, regardless of one's failure or success in the attempted learning of a second language", (Selinker, 1972, 209f.);

> "It is important to state that with the latent structure described in this paper [=interlanguage, AS] (...), there is no genetic timetable; *there is no direct counterpart to any grammatical concept such as 'universal grammar'*; there is no guarantee that the latent structure will be 'realized' into the actual structure of any natural language (i.e. there is no guarantee that attempted learning will prove successful), and there is every possibility that an overlapping exists between this latent language acquisition structure and other intellectual structures" (Selinker, 1972, 212, my emphasis).

The *interlanguage* model is a model of a latent linguistic system that is situated in a space defined by the languages in a learner's mind. It has its own systematicity and complex dynamics, and those may differ from the learners' L1 or the target language. Some 25 years later, Klein and Perdue (1997) and Klein (1998) provided another framework with the idea of *learner varieties*:

> "Learner varieties are not imperfect imitations of a "real language" – the target language – *but systems in their own right, error-free by definition,* and characterised by a particular lexical repertoire and by a particular interaction of organisational processes" (p. 538, my emphasis).

While Klein describes this to be in opposition to, or at least a radicalization of, Selinker's interlanguage concept, it is in fact not in opposition at all, but rather a decontextualization of the learner's system from their other linguistic systems (like L1 or target language ideals). The larger difference between these two, instead, is their reliance on different forces of language dynamics. Interlanguage is a latent linguistic space defined by hypotheses of the target language as made by the learner (explicitly or implicitly). As such, interlanguage in Selinker's notion is not as much an inter*mediate* (=deficient) step towards target language command, but an inter*linguistic* space with mappings from learner L1, learner target language utterances (=latent target language), and actual target language as related to by the learner. Klein's learner varieties on the other hand are not mapped onto a cross-linguistic or inter-linguistic space as defined by the three languages. Rather, they are defined by information structure and general linguistic abilities of a learner in any L2, which is also described as the Basic Variety (Klein and Perdue, 1997). However it should be noted that Klein's Basic Variety was developed on data of uninstructed SLA in an immersion context (adult migrant workers in Germany). Perhaps SLA is simply

guided by different processes at different times, more by general cognitive mechanisms at earlier stages under communicative pressure and more in relation to a (by then grown) interlanguage space later.

Both terms are frequently used in SLA research, but have never been developed to full theories of SLA. This is to a large extent due to a

> "(...) consolidation of usage-based thinking about second language learning, as reflected in the importation into SLA of an extended family of theories including emergentism, connectionism, construction grammar, cognitive linguistics, and dynamic systems and complexity theories" (Ortega, 2013, 4).

However, unlike for first language acquisition (FLA), no detailed accounts of syntactic growth from lexicosyntactic items exist for SLA. This is for several reasons: Firstly, the prototypical learner in SLA studies is the college student in a classroom setting, for the obvious reason of easy access to this cohort by researchers. There, learners are often prompted to vary and exchange material, and communicative needs require more variable language even at very early stages of SLA compared to FLA. Chunks are therefore less visible or may even move to the background of the observable. In fact, learners in an instructed setting are taught words and rules of target language generation, and typically expected to modulate their linguistic behavior in accordance with those. The conditions of study also differ from FLA in a naturalistic setting: While much of the early empirical FLA research was conducted on the children of linguists, whose parents recorded all of their speech in certain timeslots, the same is not typically possible for a learner for a number of reasons. One of those is that learners are usually not on stand-by within the reach of a linguist; another is that adult learners produce less spontaneous and repetitive speech, because they do not rely on caretakers to interact with them in the ways children do. Additionally, in a classroom, attention cannot be centered around a single speaker, so when learners rely do on different chunks, but each is only heard some of the time, overall, an image of more rule-based acquisition may emerge in the mind of the teacher or linguist.

As such, while many reasons may play a role in the lack of similar observations for SLA as FLA, there is also evidence that suggests that differences between how learners and native speakers handle chunks. This has manifold repercussions on the interlanguage or learner variety, because in FLA, chunks are learned first as communicative bits and then broken down into building blocks, but the forms of the pieces generally stay as they are. Both of this is doubted in SLA:

In their study of the teaching of chunks in a German as a Foreign Language setting, Handwerker and Madlener (2009, 12, my translation) observe that

> "[i]n contrast to the acquisition in early childhood, while school students and adults in some cases use chunks consciously for various purposes, they do not use them systematically and do not break them up as eagerly to create a base for abstractions and generalizations. They need – this is the unanimous opinion in research and teaching – a controlled intervention with explicit instruction into the usage of the prefabricated (...)".[24]

[24]The German original reads:

> "Im Unterschied zum frühkindlichen Erwerb nutzen Schüler und Erwachsene ihre Chunks zwar teilweise bewusst zu vielfältigen Zwecken, sie nutzen die Chunks aber nicht systematisch und sie brechen die Chunks nicht lernbegierig auf, um eine Basis für Abstraktionen und Generalisierungen zu schaffen. Sie brauchen – so inzwischen die wohl einhellige Meinung in Forschung

In a variation of this, Myles (2004) suggests that much of what appears like the productive use of functional syntactic categories in early SLA is in fact based on chunks, similarly to early FLA. In a study of beginning French learners in UK classrooms (years 7–9, learners were between 11 and 14 years of age) and another study of more advanced learners of French (years 9–11, ages 14–16, 'post-beginners' in her words, p. 144), she argues that

> "[e]arly productions suggested that many structures containing inflected verb forms were chunks. These structures, sometimes highly complex syntactically (e.g., in the case of interrogatives), cohabited for extended periods of time with very simple sentences, usually verbless, or when a verb was present, normally untensed" (Myles, 2004, 144).

And later (Myles, 2004, 152),

> "Chunks do not become discarded; they remain grammatically advanced until the grammar catches up, and it is this process of resolving the tension between these grammatically advanced chunks and the current grammar which drives the learning process forward.
>
> Furthermore, as learners' verb morphosyntax develops, they can be seen to add new verb chunks to their repertoire, such as *je voudrais, j'aimerais, je ne sais pas* (most advanced post-beginners only) (...)".

While this may seem like a match to FLA, it is not clear that chunks are actually ever broken down and generalized into paradigms. What Myles instead suggests that there is a re-combination of chunks into new chunks, rather than the development of generative morphosyntax. She suggests that learners

> "first map semantic representations onto phonological strings, in a somewhat approximative fashion, and in ways reminiscent of L1 children's overgeneralisations in which they pick one semantic feature of a word, e.g., shape, and overextend that word to everything sharing that feature" (Myles, 2004, 155).

Similarly, the concept of *fossilization* (Larsen-Freeman, 2006) describes a learner's coming to a halt in their acquisition of aspects of the target language before inflectional paradigms or full syntactic generativity are acquired. The result is inflectional incongruence, but it can also be understood as the chaining of unanalyzed chunks.

Yorio (1989, 62) on the other hand reports an overanalysis of chunks, like *'take advantages of', *'a friend of her', or *'are to blamed for'. Ungrammatical forms like these show that neither is the form of the chunk entirely learned nor the grammatical analysis target-like in a learner, but they also show that what is perceived as a chunk by a native speaker may not necessarily be stored as a chunk in a learner. It is also well possible that the learner picks up the chunk, but stores it not as a holistic phonetic or graphematic string, but in its analyzed form, i.e. as collocates rather than chunks. At the same time, mixed idioms like *'give up their freedom of mobility' (vs. 'give up their mobility/their freedom of movement', Yorio's example, p. 63) suggest that indeed a chunk was taken apart and part of it remapped to a similar word, which is to say that the collocates the

und Lehre – den steuernden Eingriff mit expliziter Unterweisung im Umgang mit dem Vorgefertigten (...)".

chunk is disassembled into are not necessarily of a lexical, but a conceptual kind. This is interesting, because in child language acquisition, form appears to be easy, but getting the extension of the meaning right is challenging (Naigles, 2002). For adult learners, form seems hard.

Relatedly, Stengers et al. (2011) study the the correlation between the use of formulaic language and oral proficiency in the target language in Dutch-speaking learners of English and Spanish. They conclude that "It seems that the greater incidence of morphological-inflectional errors in our participants' spoken Spanish dampens the contribution that using formulaic sequences tends to make to their oral proficiency (as perceived by our assessors)". This suggests that, unlike children in FLA, inflected words are not memorized as chunks, but paradigmatically in relation to the base lexeme, and thus do not work as facilitators of fluency or accuracy.

With this, it seems that approaches from the usage-based, emergentist, and connectionist models fail to describe chunks or coselections as they exist in L2 data with the models from L1: While adult language learners have access to generalization and specialization (constructional knowledge), they do not rely on form as much or as successfully as children do, they do not appear to be as analytical and as generalizing from the forms they do acquire, and they do not seem to be as successful at picking up conventional coselections even at very advanced levels. What then can be said about coselectional constraint in systematic perspective on learner language?

Firstly, that fixed form and coselectional constraint, although they are usually framed together in a continuum, are not generally correlated. This is not much discussed, but appears central to the differentiation between coselectional constraints and chunks in L1 and L2: Coselectional preferences can span several lexicogrammatic levels, but even on word level, they are not always contiguous. This is not much of an issue in English and French, which are the target languages most discussed in the context of L2 phraseology.[25] But in highly inflecting languages and particularly in languages with free word order, discontiguity of constituents cannot reasonably be considered 'a chunk with a slot', like in this L1 example from the Kobalt corpus (see next chapter for a detailed introduction of the corpus):

(1) Viele junge Leute können sich viel mehr leisten als frühere
Many young people can refl.pron much more afford than earlier
Generationen und trotzdem geht es ihnen dadurch nicht besser.
generations and despite_this goes it them through_that not better
'Many young people can afford much more than earlier generations, but still they do not feel better because of that' *Kobalt DEU_004*

Geht and *besser* ('feel/be' and 'better') is one of the most frequent verb + (constructional) predicate coselection in the corpus used in this study as will be shown in chapter 4. But in this example, there are four words in between those two words: *es, ihnen, dadurch,* and *nicht*, two of which could even easily be exchanged for longer, phrasal units. It is plausible that some of these are still fixed, for example *es ihnen*, *dadurch nicht*, or *nicht besser*. But in this case the frequent, segmentable, reanalyzable, and entrenched combination would not be *geht besser*, but some of the other ones.

In fact, discontiguity of associated elements is so salient in German that Mark Twain jokes about this in his essay "The Awful German Language" (Twain, 1880) at the following

[25]For some studies on L2 French, see Forsberg and Fant (2010); Myles (2004); Myles et al. (1998, 1999).

example:

> "Wenn er aber auf der Strasse der in Sammt und Seide gehüllten jetzt sehr ungenirt nach der neusten Mode gekleideten Regierungsräthin begegnet",

in his translation:

> "But when he, upon the street, the (in-satin-and-silk-covered-now-very-unconstrained-after-the-newest-fashioned-dressed) government counselor's wife met"

Auf der Straße ('in the street') and *begegnen* ('meet') in coselection are a conventional way of saying 'to randomly meet someone' in German (366 hits in the German reference corpus DeReKo (Leibniz-Institut für Deutsche Sprache, 2019) for the identical *auf der Straße begegnen* without extra words, paradigmatic occurrences uncounted), but are 14 tokens apart in Twain's example.

Secondly, both fixed language (like chunks) and coselectional constraint are functional in language, and both must be acquired. In a systematic model of interlanguage at least in learners in an instructed setting, chunks, words, and rules compete in early acquisition stages. However, learners also know few words and do not have much to coselect. At the same time, each coselection of two words reinforces the connection. With growing skill, more words and syntactic constructions can be coselected, but since constraints are opaque to the learner, they cannot be adhered to. This must be expressed in a process of differentiation (more words) and randomization (random, non-nativelike coselections). With some implicit and some explicit learning, and perhaps also through the strengthening of systematic properties of the representation of the target language in the learner's mind (semantic differentiation, phonotactic aspects), constraints are then re-learned in a process of specialization. However since this process reflects only a skeletal layer of the amount and the development of input processing in L1, and native-like levels are rarely reached. This process would be reflected in a u-shaped development, not unlike the one that was described for L1 in earlier sections of this chapter. But since L1 and adult L2 acquisition differ by many functional and contextual factors, while both developments are u-shaped, and both are built from chunks, to free combinations, to coselectional restrictions, it is to be expected that learners and native speakers develop those on different categories and exemplars and with different ease and timing. In fact, u-shaped developments are simply models of dynamic re-organization, where rules once formed lose efficacy at one point in the development and thus require rearrangement and reformulation. Carlucci and Case (2013) formulate u-shaped developments as a necessity for any process that includes general tendencies and idiosyncratic or item-based specifics.

Thirdly, some research suggests that target language immersion is of particular benefit for the acquisition of formulaic language in learners, more so than for the syntactic or lexical development (Foster et al., 2014). While part of this may be a reflection of higher exposure (frequency and memorization), the fact that the same degree of a benefit does not exist for the syntactic or lexical development suggests that coselectional constraints, at least on a word level, may be enculturated and semiotic rather than arbitrary. In fact, one aspect that majorly differs between learners in an immersive setting and outside one is contextualized communication with native speakers. Learners in a classroom setting may lack a functional necessity for the acquisition of coselectional constraints, because even when they speak with their classmates in the target language, the semiotic system is still the L1 context.

2.2.4. Summary

In usage-based approaches, language learning is modeled as an emergentist and connectionist process, where generalizations over frequency distributions lead to higher level generalizations and thus the formation of a functional grammar. However it is rather unclear how coselectional constraints, for word level coselections in particular, are learned in this process. Neither the model of segmentation of phonetic strings nor distributional learning seem to fully match the problem that is sought to explain, because some coselections are discontiguous or span several levels of lexicosyntax, thus being unavailable or at least unlikely candidates for phonetic segmentation, and because some are so rare that the distribution of a category never realistically unfolds in natural input. In addition, models of FLA fail to describe data from SLA where the phonetic form seems less salient and memorable to learners, and chunks are broken down, but often for rechunking and chaining rather than generalization.[26] Furthermore, the functional or interactional benefit of chunks as direct communicative bits is not typically a characteristic of coselectional preferences.

What is known from SLA research is that learners have difficulty acquiring and using coselectional preferences in the form of collocations, where early learned chunks seem to persist ("phrasal teddy bears") and be used disproportionately, but the overall number of collocations that is used or understood, even in the frequency range of the 1000 most frequent collocaitons, is low. Even advanced learners struggle with this. This is generally ascribed to the idiosyncratic, unpredictable nature of coselectional preferences. Indeed, some studies from category learning in L1 have shown a sensitivity to narrow-ranged semantic features and phonotactic characteristics of lexemes in the coselection with syntactic slots in adults, but not in younger children. It seems then that adults are sensitive to this aspect of language. It is also possible that what has been observed as idiosyncratic (see section 2.1.3) is not in fact idiosyncratic, but follows intricate rules rooted in more-fine grained distinctions than are currently made in the models. However, it is unclear why L2 learners should not be sensitivite to the same aspects, since it can be assumed that they exist in their L1, too.

One possible explanation is a lack of semiotic mapping between concepts and their expression in coselectional constraints of the target language. Learners, unaware of certain signs that are coined in the target language, instead reassemble them from scratch, which may never quite fit the exact semantic shape of a convention in the target language. This may be what makes an expression sound unidiomatic. It might then also be that coselectional constraints are not learned by word or construction, but by communicative or semiotic situation, i.e. as a category of a sign with prototypical linguistic expressions of the same. A semiotic perspective would explain the high rate of L1-transfer in the use of collocations in learners and the strong benefit of even short-term enculturation, which does not seem to occur equally for other, and more frequent, aspects of the target language system.

2.3. A theory of coselectional constraint?

So far, this chapter has shown that there is a tendency of words and lexicosyntactic and syntactic constructions to co-occur with one another in preferential ways. It is unknown as of yet how stable and specific such forces of attraction are on a lexicosyntactic level,

[26] The role of a potential implicit reorganization in the abscence of evidence cannot be discussed here.

and whether they are guided by fine-grained semantics, morphosyntactic or phonotactic regularities, or are truly arbitrary and idiosyncratic. But the observation that the full productivity and combinatorial power of syntactic, lexicosyntactic, and lexical elements is not realized in language in use, as it is conceptualized as a general tendency of developed native speaker language by Firth (1957), Sinclair (1991), and Pawley and Syder (1983), has been solidly confirmed in many studies.

In the theoretical model of usage-based linguistics, however, this has not been integrated in a comprehensive way yet. Coselectional constraints are instead frequently treated as belonging to a continuum from fixed to freely combined language, i.e. as partially fixed constructions. The continuum model was introduced around the same time for phraseology by Bahns (1993) and Howarth (1998) and for grammar as a whole by Goldberg (1995) and is unanimously accepted in the field:[27]

> "Idioms (...) are relatively frozen expressions whose meanings do not reflect the meanings of their component parts. An example containing the noun murder would be *to scream blue murder* ('to complain very loudly')- Between idioms and free combinations are loosely fixed combinations (or collocations) of the type *to commit murder*. (...) There are, however, 'transitional areas' (...) between free combinations/collocations and collocations/idioms" (Bahns, 1993, 57);

> "(...) it is unlikely that grammar consists of a set of productive rules, a lexicon and a collection of frozen phrasal idioms. Instead, these 'modules' are permeable. (...) [C]onstructionist theories make this interaction particularly seamless by providing a single representational format for productive processes, tightly bound idioms, *and everything in between*" (Michaelis, 2012, 57, my emphasis);

> "Formulaic language can be of many different kinds, such as, collocations (*fast food*), binomials (*black and white*), multi-word verbs (*rely on*), idioms (*tie the knot*), speech formulae (*what's up?*), discourse markers (*by the way*), lexical bundles (*as well as*), expletives (*damn it!*), grammatical constructions (*the –er the –er*), and many more" (Siyanova-Chanturia, 2015, 286).

There is indeed one functional aspect aspect of recurrent speech that is shared by chunks and coselectional constraints, which is the limitation of the semantic search space in meaning negotiation between two speakers. Wray (2002) proposes this as the main function of formulaicity in language. However, there is a number of important differences between chunks and coselectional constraints:

1. Chunks are form-meaning pairs (one form, one meaning), while many coselections exist in a morphosyntactic paradigm where several forms are mapped to the same meaning, most notably in highly inflecting or agglutinating languages;

2. Chunks are continuous phonetic or graphematic strings that are available for analysis and segmentation, while collocations are often discontiguous and may span distances

[27] A more radical variation of this approach are lexicalist grammars like Pattern Grammar (Hunston, 2012), Word Grammar (Hudson and Hudson, 2007), or purely connectionist word priming grammar models (Hoey, 2012). In those, all linguistic abstractions are viewed as epiphenomenal, and all language basically consists of chunks with slots. However it has been argued that for languages with flexible word order, this does not appear highly plausible. See Müller and Wechsler (2014) for a comparison and critcism of lexicalist grammar models.

that cannot be bridged by the working memory (like very long utterances or sentences in German, where a verb occurs in the finite position far, far away from its object);

3. Chunks can cross syntactic boundaries, like *is the*, while coselectional constraints are category-bound (verb + object or adjective + noun rather than word + word);

4. Chunks can be chained, coselectional preferences cannot;[28]

5. Chunks are easy to learn for learners, while collocations are hard, at least in general terms. How this applies to individual coselections or chunks cannot be said with certainty from the current state of research;

6. Chunks are learned first in FLA and then broken down and rearragned, co-selectional preferences likely remain underdeveloped until late school age, there is also likely a larger variation in coselectional prefences than in chunks in L1;

7. Chunks are usually described as frequent in the literature, coselectional preferences exist in all frequency ranges;

8. Chunks are so easy to memorize that a number of non-human animals can learn to either recognize them (cats, dogs) or produce them (a number of birds, like crows and parrots); Coselections require syntactic embedding and are therefore not usually learned by non-human animals;

9. Chunks support fluency because they do not require full morphosyntactic processing (they are essentially processed like words some of which have gaps), while coselectional preferences do, and their influence on fluency is unknown;

10. Chunks remain productive in speakers with aphasia or dementia, while coselectional preferences are unstudied in the context of language impairment but are unlikely to have the same effect;

11. Chunks are ready-made for social and cultural use, while coselections always require further linguistic context and embedding.

This list is likely not comprehensive, but it already shows that in at least three main branches of theorizing: formalization or classification, function, and development, differences do not appear to be of a gradual kind. Importantly, chunks, if they are analyzed, are of course instances of coselection, in the same way that chunks are frozen instantiations of syntax, morphology, etc. The problem is not that chunks can be considered coselections (depending on the degree of analysis that is assumed), but rather that there are coselections that are not very chunk-like, and lying at the very interface of lexis and syntax, and of *langue* and *parole*, underly much more complexly interwoven and dynamic processes than chunks do; and that this is essentially unmodeled at present.

2.3.1. Conflations in the continuum hypothesis

What then leads to a unified treatment of such different phenomena? Part of the explanation is likely the historical development of usage-based and functional grammar,

[28] This is because coselectional preferences cross category boundaries and require syntactic embedding. Of course *coselections*, i.e. realizations of coselectional preferences or constraints can be chained.

where the idea of a continuum from lexis to grammar grew in opposition to the existing sharp division of the two viewed as separate modules in the generative paradigm.[29] As such, cognitive-linguistic models and construction grammar (CxG)[30] in particular not only assigned semantic meaning to all syntactic constructions, but grammatical aspects (like productivity) to all lexical items, too. However, if the lack of differentiation between coselectional constraints and chunks was mainly historical, it could easily be resolved by integrating a separate process. Another reason is likely the focus of empirical linguistics on more analytical languages with a fixed word order like English, French, or Mandarin, which may lead one to believe that all coselections are in fact chunks with fewer or more slots. But it appears that there is also a deeper problem that lies in a lack of differentiation of a number of concepts.

The quote from the beginning of the first chapter by Yorio (1989), where some language is described as being *more* conventionalized than language already is, shows that conventionality and conventionalization denote a complex concept that requires detailed analysis. Convention has always been a topical theme in linguistics since Saussure (1916/1983) described language as a system of arbitrary, conventional form-meaning pairs called signs. Wittgenstein (1953) in his Philosophical Investigation defines the meaning of a word through its usage (*Gebrauchstheorie der Bedeutung*, 'usage theory of meaning'):

> 'One may, for a large class of cases of the use of the word 'meaning' – even if not for all the cases of its use – explain the word in this way: The meaning of a word is its usage in language. And the meaning of a name can sometimes be explained by pointing at its bearer' (Wittgenstein, 1953, §43, my translation).[31]

Usage-based linguistics combines the two by saying that language is a set of linguistic signs or constructs that are form-meaning pairs (Saussure) and those are created from experience with the language in use (Wittgenstein). Through this connection to habitual usage, conventionality implies a certain 'sameness' of things that appears similar to the concept of fixedness, and usage also appears to imply frequency; Conventionality in Saussure's lectures was also equated with arbitrariness, which may be understood as idiosyncrasy (more will be said shortly about this equation). With this, convention is modeled as a major force of constraining the otherwise uncontrolled productivity of lexicon and syntax:

> "A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" (Erman and Warren, 2000, 31).

[29] It should be noted that factually those models still divide between lexicon and syntax, because many construction slots only accept lexemes as fillers, and because a form-meaning pair requires a form, which a fully abstract construction does not have (it has form*s* that it is realized as). See also Boas (2008a,b).

[30] For some models of CxG, see Sag et al. (2012); Sag (2012); Boas (2013); Goldberg (1995, 2006); Croft (2001). While not all of them use the abbreviation CxG, it will be used without further differentiation here.

[31] The German original reads as follows:

> "Man kann für eine große Klasse von Fällen der Benützung des Wortes »Bedeutung« – wenn auch nicht für alle Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache. Und die Bedeutung eines Namens erklärt man manchmal dadurch, daß man auf seinen Träger zeigt".

For a critical perspective on whether this can be meaningfully related to the meaning of usage in usage-based linguistics or to linguistics in general, see Thiele (1990).

> "Moreover there has to be a way to coordinate the unavoidable variation in language use so that a shared set of conventions arises and is maintained by the population, even though there is no central coordinator, nor a prior innate grammar or telepathy" (Steels, 2013).

Even in Wulff (2008)'s profound account of idiomaticity as a multidimensional phenomenon, the aspects of fixedness and non-compositionality are predictive of the rating of idiomaticity by native speakers (students of English), but corpus frequency is not:[32]

> "The present study is the first to present an approach that comprises both semantic and syntactic variation and assesses the relative importance of each variable. The results tie in very well with many widely established claims about idiomaticity, with tree-syntactic flexibility, particularly passivizability, as a key characteristic to describe the distribution of V-NP-constructions. Likewise, the central role of compositionality is reproduced by the PCA [principal component analysis, AS]. However, the multifactorial perspective reveals that when being considered *in toto*, other variation parameters turn out to be even more important, namely aspects of morphological and lexico-syntactic flexibility" (Wulff, 2008, 164).

In this conclusion, conventionality outside of morphological, semantic, or syntactic inflexibility is not included, yet still she agrees with the continuum model:

> "(...) descriptions of the constructicon mainly focus on the gradience regarding the lexical specification and structural complexity of constructions. This creates a continuum of constructions in which idiomatic expressions of the kind analysed here are located somewhere in the middle" (Wulff, 2008, 164).

However, convention, fixedness, frequency, idiosyncrasy or arbitrariness, and forces of attraction are not all the same phenomenon, and their overlap is only partial at best:

[32] Wulff collects native speaker judgments of the idiomaticity of 39 verb + noun coselections (she calls them constructions) of various kinds. Some are syntactically flexible and semantically compositional but conventionally co-occurring, like *tell + story*, and some more opaque and inflexible, like *bear + fruit* or *take + plunge*. In her model of contributing factors to judgments of higher idiomaticity, the parameters chosen account for 0.565 of the variance (adjusted R^2), which is not overwhelming (regular R^2=0.794, but with 20 parameters for 39 coselections, hence the low adjusted R^2). This is not an overwhelming correlation, but solidly suggests that native speaker intuitions correlate with linguistic analyses. She summarizes a factor-based analysis as follows, where beta weight indicates an estimation of how much each parameter contributed to the regression in the range [0,1]:

> "(...) [T]he most important variation parameters are the morphological flexibility parameters NumV and Mood. [at a beta weight of 0.757 and 0.695, AS] They are followed in rank by two lexico-syntactic flexibility parameters, KindAdv and NoAdv (...) [0.651, 0.632, AS]. Next in line are compositionality and tree-syntactic flexibility. The morphological flexibility parameters Voice and Neg also yield sufficiently high beta weights to be considered relevant. The last variation parameter with a value higher than +0.22 is the lexico-syntactic flexibility parameter Addition (...). Corpus frequency (CorpFreq) yields a beta weight of only 0.209" (Wulff, 2008, 159).

The most idiomatically ranked coselections are *foot + bill*, *meet + eye*, and *bear fruit*; the least idiomatically ranked *write + letter*, *tell + story*, and *call + police* Wulff sets a cut-off point at $\beta \geq 0.22$, which indicates that the parameter accounts for five per cent of the variance. This is just above the variable of corpus frequency. All higher ranked aspects are subcategories of morphosyntactic inflexibility and non-compositionality.

- A thing is fixed if it cannot easily or without loss be changed in a certain respect, or if it could be changed but is not.
- A thing is conventional if it is habitually used or done in a specific way in a given context.
- Conventionality is re-established with each use, while fixedness requires no re-establishment since its wholeness is intrinsic to the item. For example, I may reinforce my personal convention of wearing a helmet for riding my bike, but I need not reassemble the bike. Convention also has a higher-order function: Rather than for keeping together helmet and head per se, it is supposed to serve safety (I could not wear the helmet and accept a higher risk). If the fixedness of my bike is lost, its function leaves with it completely, while both my head and the helmet keep their function if I decide to keep them apart.
- An item or a coselection is idiosyncratic if it has a feature that is unique to its mechanism and cannot be extended by or derived from analogy.
- It is arbitrary if there is no outside force that could explain the idiosyncrasy. For example, the Basque language is idiosyncratic in a modern European context, but its existence, structure and development can still be understood through historical and linguistic study (provided the necessary sources), it is not arbitrary.
- A force of attraction is a tendency guided by an outside cause that materializes between specified elements under friendly environmental conditions, like two magnets will only attract each other if there is no stronger magnet around, if they are close enough and facing each other. It is not the same as an intentionality. There is no specific force of attraction between my bike helmet and my head, the two are brought together by my intention only. But there appear specific forces of attraction between the cognates of two languages in the second language learners mind (Rabinovich et al., 2018; Prior et al., 2007). Thus, some of the concepts named are even mutually exclusive, like fixedness and forces of attraction (what is fixed requires no force of attraction to keep it together).

None of these concepts entail high frequency of co-occurrence per se: There is a conventional way of celebrating a wedding, but it is not typically a frequent occurrence in the life of an individual, and not necessarily celebrated in a fully fixed and ritualized way either. In fact, the more conventional a person chooses to be in that respect in Christian (and many other) societies, the less frequent their weddings become. Of course in a society as a whole, convention and frequency may be somewhat correlated, but there are also conventions that materialize only very rarely. To give a few linguistic examples:

- A chunk is fixed, because if it is teased apart, it stops being a chunk and loses its processing advantages as they were described earlier; its parts do not require forces of attraction, because they are not analyzed as parts.
- An idiom is idiosyncratic, because its meaning cannot be derived from the rules of semantic composition and the latent regularities of its meaning cannot be extended to other idioms (*to bite (the dust)* or *to kick (the bucket)* may mean *to die*, but this cannot be extended to *to draw* in *to draw the line*). It needs not be fully fixed, but if an essential part of its structure is taken away, it loses its unique status (as in

she kicked several buckets). If a collocation like *make amends* is used outside of the formally expected, as in **she make amends*, its function will likely persist. An idiom is not necessarily arbitrary either, many idioms are in fact obvious metaphors or easily explained through historical context. Some collocations, on the other hand, appear to be more arbitrarily restricted, like *make a list, make amends, make a choice, make profit*, but not **make an experience* (which in German is not only not blocked, but the idiomatic way to express 'to gain experience' or 'to experience': *eine Erfahrung machen*).

- Social routines are conventional, they exist and are used for a higher purpose, but are neither fixed nor idiosyncratic. For example greetings can be replaced by other specialized routines like secret handshakes (which are then idiosyncratic). Forces of attraction can be semantic, for example, as suggested by frame semantics where a certain frame evokes a number of agents, processes, and objects which are bound by semantic cohesion; and expanded upon by distributional semantics, where the forces of attraction between words *are* the semantics.

- An item may be highly idiosyncratic, like a proverb or an idiom with a very specific meaning, and it may be frequent or infrequent. A chunk may be frequent (like *I think*) or infrequent (like many proverbs, see Moon (1999), or even existing word forms that are listed in the Oxford dictionary but occur only once in the British National Corpus (Deshors et al. (2016), meaning also that they occur only in one coselection of each lexicoyntactic type).

- Coselections may be idiosyncratic, like highly specialized vocabulary (*tokenize a corpus*). Tokenization is a process idiosyncratic to text, it cannot be extended to other entities, but the form of the coselection is not fixed at all (some possible modifications include TAM, pluralization, exchanging 'corpus' for a corpus name).

- A ship's christening, one of the prime examples of Austin's speech act theory (Austin, 1975), is a highly conventionalized, and a rather fixed, but not a frequent event compared to other events like answering the phone or speaking to colleagues; It is not idiosyncratic (there are other types of christenings), and it is hard to argue that there are specific forces of attraction between the linguistic items or the people and processes involved that would explain the result.

- Wishing someone a happy birthday in German, on the other hand, is a conventional and frequent, but not a highly fixed process, there are several ways to express good wishes[33] and a number of songs that are frequently and interchangeably sung to the occasion.[34] Depending on the kind of party, idiosyncrasies may occur (like humorous poems written for a special birthday of which the specific form may be idiosyncratic); and semantic forces of attraction may play a role in all related linguistic aspects, namely through evoking a celebratory frame.

Of course what is fixed can only be used in its fixed way. So if the meaning is bound to a fixed form, an item can only be used in that form, forcing conventionality onto fixedness.

[33] For example *Alles Gute; Alles Liebe; Zum Geburtstag viel Glück; Herzlichen Glückwunsch; Happy Birthday; Die besten Wünsche; Glückwünsche; Viel Glück und viel Segen; Glück und Gesundheit; Von Herzen nur das Beste.* Of course some of these are more prototypical and frequent than others, but the first four + the one borrowed from English are very common and interchangeable.

[34] *Happy Birthday/Zum Geburtstag viel Glück; Hoch soll er leben; Viel Glück und viel Segen; Wie schön, dass du geboren bist.*

And on the other hand, fixedness may grow from conventionality through entrenchment. Yet it may be that some items are both fixed and conventional in a given context, while the fixed item may be unconventional ('unidiomatic') in another context, without ceasing to be fixed.

This is not the place to expand on all dimensions and their combinations, but let it be said that with six dimensions (convention, fixedness, arbitrariness, frequency, forces of attraction, idiosyncrasy), a simplified model that expects no mutual exclusivity amounts to $2^6 = 64$ possible combinations. Even if half of those were impossible to attest with linguistic examples (which I doubt), this would still leave 32 combinations including fixedness as a factor, or 16 combinations that are not fixed at all. This is also reflected, but not reflected upon, in the study of collocations and formulaic languages with its many terms and subclasses.

Thus, while it may not be wrong to arrange chunks, collocations, coselectional constraints, etc. on a continuum from more to less fixed, the explanatory power of this model is low:

- It has low classificatory efficiency – items cannot be arranged by idiomaticity through fixedness alone and the other dimensions are not clearly definable as correlations of the degree of fixedness in a way that idiosyncrasy or non-compositionality could be modeled as covering certain parts of the continuum;
- It is not a good developmental model of L1 or L2, because coselectional constraints are located at intermediate levels in the continuum, but are acquired last;
- It is not a good model of native speaker intuitions: While morphosyntactic fixedness is a good predictor of ratings of idiomaticity by native speakers (Wulff, 2008), corpus frequency without fixedness or non-compositionality is not, suggesting the two are not well correlated in perception;
- The function and the linguistic realization appear to be dichotomous rather than gradually increasing, since coselections must always be realized in a lexicosyntactic context:
  - Morphosyntactic processing is necessary in coselections, while chunks cannot be integrated in a different linguistic context and are not processed morphosyntactically; whether or not morphosyntactic processing is activated is dichotomous (even if it was always activated additionally for chunks in speakers without language impairments, for those with language impairments the dichotomy becomes visible);
  - Whether or not something is a communicative bit is dichotomous. Not all coselections are *not* communicative bits, but some are, like *trick into believing*; or *give* + DITRANSITIVE CONSTRUCTION, or *primary aim*.

With this, the model offers little prediction and systematization of coselectional constraints. What seems to be in order instead is a clarification of the role of frequency, convention, and narrow-range semantic and morphophonotactic regularity (Wonnacott et al., 2017; Ambridge et al., 2012, 2014), and true idiosyncrasy and arbitrariness.

2.3.2. Convention in usage-based grammar

It has already been said that convention is a complex phenomenon. The general agreement in usage-based grammar is is that constructions are lexicogrammtical signs, i.e. form-meaning pairs. While his is close to Saussure's definition of a sign, it is unclear that they truly are: There are sets of meanings that can be mapped to a common form, like homonyms or garden-path sentences; and there are sets of forms mapped to the same meaning, like morphosyntactic paradigms or near-synonyms (Imo, 2011). This suggests that meaning is not directly attached to the form, but rather that the form and the referent of a sign map to a common concept, and this concept may be attached to other forms as well. This is a triangular semiotic model like the one suggested by Ogden and Richards (1923).

In both models (pairings vs. triangles), the question remains whether coselections have *one* meaning, i.e. whether they are stored as holistic items with a common meaning (and in whom); or whether they are connected only statistically in corpora, but not actually mapped to a single meaning; and what guides their coselection if they are neither frequent nor fixed. Unlike in a fully fixed construction, a semantic merging must occur at some point for all constructions involved in the generative production of an utterance (the syntactic construction, the lexeme, the larger linguistic context like register, etc.). How is convention involved if coselections create a merged meaning, and how is it if they do not? Are coselections productive, or are they fixed in meaning even where they are not fixed in form?

Since coselections are conceptualized as a type of chunk, they go essentially unmodeled in CxG. Where they are part of a partially filled construction, the slots may specify a slot-filler lexeme, although it is unclear why they would then not be stored as a chunk – unless it is not *one*, but a distribution of lexemes that is specified by the slot. This however is an odd mixture of abstract and concrete aspects within the same construction that is currently not part of the model to my knowledge, and it is unclear which form to attach the meaning to, and what the specific meaning of the coselectional construction and each of the lexemes is. Where they are not part of a partially filled construction, it is unclear where to model coselectional preferences in CxG. Of course one way would be to define each word as a coselectional construction where the slot is lexically specified. As the most formal of construction grammars, sign-based construction grammar (SBCG), has been suggested as nearly identical to head-driven phrase structure grammar (HPSG) (Müller, 2017), but it is particularly the long-distance dependency of some coselections that is difficulat to model. Since transformations are generally not welcome to CxG, each transformation (or use in a separate syntactic construction) that *is* allowed for an unfixed, but conventional coselection, exists as a separate construction but exerts the same force of attraction on a lexeme (see Sag (2012) for a summary of SBCG).

In HPSG, a few suggestions have been made by Erbach and Krenn (1993); Richter and Sailer (2009); Cook (2014), where phrasal elements (chunks with slots) are modeled with a semantically specified slot, idioms are listed under a separate word sense, and a LEXEME-feature has been suggested for coselectional constraints on word level (collocations). It is unclear whether the whole distribution of coselectional preferences should be included under this feature, and whether coselectional preferences should be modeled as mutual (noted in both lexeme's signatures), or how to quantify them. Listing *all* coselectional preferences in *all* signatures seems like a huge redundancy (this may or may not pose problems depending on the presumed model of memory) and would also require for signatures to be able to cross over to other signatures and change their coselectional preferences: If

the association strength between two coselected items is weakened, the relative association strength of other items may grow. This can be modeled in a graph (see chapters 5 and 6), but it is not entirely clear how to fuse this with different levels of lexicosyntactic granularity.

Of course all of these questions are not merely shortcomings of the grammar models themselves, but are unanswered in all of linguistics currently: It is unclear whether convention is attached to lexemes and lexicosyntactic constructions separately or to their combination, i.e. whether words and constructions coselect directionally (one coselects the other), or if they coselect each other mutually in a single choice (as Sinclair (1991) suggests).

What is typically used as a measure for deciding whether coselections are holistic or recombined is their frequency of occurrence. But this is misleading: It has been mentioned several times now that idioms occur rarely in corpora. But moreover, as Moon (1999) states, some of the most fixed phrasal idioms occur only rarely *in their natural form* in corpora. Instead, they are often altered in an act of discourse reference to a known prototype. But if the prototype can be refered to through analogy, then this means that forms that are not the form of the sign itself can activate the meaning, i.e. access the concept from outside of the semiotic triangle. However, this also means that it is rather unclear how to measure frequencies adequately. If a fixed form can be activated (and hence entrenched?) not only by occurrences of itself, but also of similar items, how does this translate to coselections, and how can this be constrained by conventionality? It appears in fact that construction grammar and lexicalist approaches moved part of the complexity that they sought out to explain out of the way by defining it as intrinsic: Things are recurrent in language because they are conventional, and they are conventional because they are recurrent. But even those that are not frequently recurrent still may be judged as conventional:

> "Just as some rare but conventional forms of British English appear only once (if at all) in the British component of the International Corpus of English (ICE-GB) (...) or the British National Corpus (BNC 2007) (in morphology, for instance, a number of words ending with the suffix –ness, such as overtness or effortlessness, are hapax legomena in the BNC, but are recorded in the Oxford English Dictionary (OED 2015) and are thus conventional forms)" (Deshors et al., 2016, 10 in preprint).

In summary, conventional coselection cannot be explained or modeled adequately with the existing principles of usage-based grammar. In fact, Dux (2016, 427) points out that the observation of idiosyncrasies in the coselection of verbs and argument structures is not only problematic for projectionist approaches, but

> "equally problematic for constructional approaches such as that of Goldberg (1995, 2006), whose principles state that verbs may be used within a given construction if the verb's participant roles are semantically compatible with the constructional slots of the construction. Given the semantic similarity of Change verbs (and thus their participant roles), one would expect that these verbs would be equally felicitous in the same range of constructions. Again, this conflicts with the data discussed here, necessitating a reformulation of the principles for verb-construction fusion".

Goldbergian CxG does not view itself as a constraint grammar and would not be bothered by the theoretical overgeneration, but rather rely on conventionality to constrain

those preferences.[35] This is a way of shifting the complexity of constraint grammars to an arbitrariness implied in convention, but as has been said earlier, convention is not necessarily arbitrary. While some conventions can only be explained causally from a historical analysis, this does not mean they serve no function in the present. For example, some religious ceremonies or festivities that are kept up by atheists, like Christmas, may not serve the original religious purpose, but still that does not mean that they are meaningless or purposeless. Neither does it mean that they are not anchored in specific, systematic ways that have repercussions for the organization of the mind or social systems.

This is not to say that convention plays no role in coselectional preference, it probably explains a significant proportion of the data. But one aspect of convention is that one must know it to be able to adhere to it, which implies also that a conventional way of expressing a certain meaning exists in the language. In the case of coselectional preferences, a native speaker may recognize what a learner was trying to say and correct them to the conventional sign. The idea with this is "oh, I understand, but this is not how you say it – we say it like that: ...". There is another case though that fits an earlier example from section 2.1.1, where a German school student in 8th grade was reported to say *um ihre Höherachtung zu bekommen* 'to gain their higher-respect'. For both the author of the referred paper (Hee, 2019) and for me as a native speaker of German, this is odd and unidiomatic, which is to say *it sounds wrong*, but I cannot think of a conventional way of expressing this meaning either. Perhaps then there is even a difference between conventional coselection (=coselectional preferences) and coselection constraints. Most conventional coselection would likely be constrained by coselectional constraints, but not all coselectional constraints would be manifested in a conventionalized coselection.

2.3.3. A research agenda for understanding coselectional constraint

Many questions have been raised in this section regarding the right formal definition, the function, and the cognitive and social constraints that influence coselectional preferences and constraint. As Tucker and Fawcett (1996, 147) point out,

> "[c]ollocation is not, in itself, a theory of lexis. Yet, as a relation holding between lexical items, in terms of co-occurrence and mutual selectivity, it must be incorporated in any theoretical account of lexis, and therefore in any theory of language".

The same is true of coselection on all lexicosyntactic levels. In this spirit, a formal integration of coselectional constraint into usage-based grammar, beyond the continuum hypothesis, should be aspired to. One of the challenges of such an integration lies in the systematization of the dimensions as they were named in section 2.3.1, and to learn more about the mutual vs. the unidirectional coselection of items, their structural role or repercussions, and how the combinatorial power and potential redundancy of this can be managed in a formal and a cognitive model. Another is to find an answer to the question

[35] There are other construction grammars which are more constraint-based, like Sign-Based (Sag, 2012) or Fluid Construction Grammar Steels (2013), although it appears that a latent contradiction of constraints vs. acceptance of idiosyncrasy is inherent in the model of constructions. This can be resolved in the dynamic perspective, i.e. for how constructions are learned through dialectic interaction of the two principles, or how they change. But it appears difficult to predict generativity in synchronic or static use, i.e. to draw the line between what should work, what does work, what could work, and what is not used after all etc. in construction grammars (unless they are very close to HPSG, see Müller (2013b)).

of whether convention is an a-priori force, what its relationship to frequency and entrenchment is, and whether and if so where a line can be drawn between narrow-ranged semantic and morphophonotactic regularities and true idiosyncrasy or arbitariness. Related to this is the necessity of a functional description of coselectional constraint.

There are some implications in usage-based accounts in this regard, spanning two polar ends of a scale of relevance: From a semiotic account by Wray (2002) on one end, who suggests that using the same words for the same contexts limits the semantic search space by priming specific contexts – which should then extend also to coselectional preferences – to emergentist approaches where 'what is used together, fuses together' (Bybee, 2002), such that coselectional preferences may be interpreted as a statistical epiphenomenon (it is the frequent use that explains the constraint, not the constraint that explains the frequent use).

The history of linguistic research shows that what is modeled as an epiphenomenon of one category often turns out to in fact not be fully explained by or even well correlated with it – like verb distributions across verb-argument structures that were once modeled as epiphenomena of semantic congruence and turned out to have an idiosyncratic tendency; And it may in fact function in its own right and with its own purpose – like it is being discussed for disfluencies and hesitation phenomena that were initially analyzed as processing deficiencies (Tottie, 2011; Kidd et al., 2011; Belz, submitted).

It is of course possible that there is nothing to find, that convention is randomly sprinkled over language and is, albeit wide-spread, of peripheral interest to the linguistic study. But the same was said of formulaic language at first, and later it turned out that chunks play a role in fluency, language learning and change, negotiation of meaning and of course in any number of social routines (Pawley et al., 2007), not only in speakers whose generative skill is not yet or not anymore at full capacity. It would be wise to assume that if conventional coselection is observable between linguistic items on different levels of granularity and cannot be explained well by the existing theories, that this is not merely a random or statistical effect, but a functional aspect of language that is not well understood as of yet.

> "The first aim of scientific research cannot be the mere accumulation of knowledge. It must try to uncover the general principles behind the masses of particular findings: It must eventually come up with a "theory"" (Klein, 1991, 49).

Yet before it can do so, it must first find an analytical frame that allows for meaningful comparison, because only findings interpreted in a common analytical frame might eventually converge into a unified theory. For this, a systematic study of coselections is needed. Since little is known about coselectional preferences in the use of individual speakers or homogeneous cohorts of speakers in homogeneous texts, my approach in this thesis is one that considers both L2 and L1, and L2 as a dynamic process. This will be approximated here with a cross-sectional study of Belarusian and Chinese learners of German as ranked by standardized test scores, who will be contrasted with native speakers of German, where all data was intended to be as homogeneous as possible. This allows for a comparison of both L1 and L2, with learners from two different L1s, and a comparison of learners at different stages of acquisition. The idea is that if there are dynamic processes, they should become visible in learner trajectories over acquisition stages. If there are structural properties implied, they should be observable in the most similar ways in very advanced learners compared to intermediate or beginning learners; and they should be most observable between most advaced learners across language groups, because they are

closer to the target language space, while intermediate learners might be closer to their respective language groups. If furthermore coselectional constraints or preferences play a structural role or at least bear structural consequences, then learning trajectories should follow certain rules of reorganization, and not be erratic. A set of specific hypotheses will be developed in the next chapter.

While the dataset used marks only a very specific register of L2 and L1 German, and while it is relatively small, I believe the study still provides new material to some of the questions into the study of coselection that have been implied:

- whether fixed chunks are to be interpreted in the same way as coselectional preferences;
- how coselectional preferences are represented in texts quantiatively and qualitatively (how many, what kinds are there, are they all the same?);
- whether (and how) coselectional preferences are different in learners and native speakers, both quantiatively and qualitatively;
- what can be said about the *structural* role of coselectional preferences or constraints in SLA

These will be discussed to varying degrees throughout the thesis and addressed with new insights from this work in chapter 7.

3. Hypotheses and data

In this chapter, hypotheses for the corpus study that is at the core of the thesis will be derived from the theoretical background discussed (section 3.1). The data, a mid-sized corpus of essays written by German SLA learners at Chinese and Belarusian universities, compiled by the Kobalt project (Zinsmeister et al., 2012) and processed by the Kobalt project and myself, will be presented in section 3.2. This includes a discussion of the value of the simple, but validated onDaF test (Eckes, 2017) as a ranking or grouping variable (section 3.2.1) and of some of the annotation choices that were made for the specific purpose of the study in section 3.2.2. Annotations that primarily concern the graph-based model will be reported in chapter 5.

3.1. Hypotheses

To recapitulate and conclude from the previous chapter:

- Constraints or preferences in the coselection of lexicosyntactic structures appear to be a property of natural language. Although to date there exists no precise linguistic model that would characterize and systematize coselectional constraint, one observation is recurrent: Idiomatic or 'L1-natural' combinations tend to be elusive to derivation by semantic or other obvious features in most cases. This cannot be said with absolute certainty, because it remains possible that a more fine-grained semantic categorization or morphophonotactic rules or tendency guide the process. From what has been modeled, however, coselectional constraint appears to reflect properties of individual items or clusters of items rather than cross-system rules. This has often been modeled as a force of attraction an item has on another or as the conditional probability of the two occurring together as opposed to separately.

- Coselectional constraint thus is an intricate and likely partially idiosyncratic phenomenon that needs to be acquired in both L1 and L2.

- It may not be possible to formulate generalized rules predicting which coselections would be acceptable. Despite this, on a more abstract level, coselectional constraint is also a structural property of language: On the whole, language can be used more formulaically or more productively, corresponding to Sinclair's *idiom principle* and *open choice principle* (Sinclair, 1991). Structural properties of a language would be reflected in the structure and composition of the mental lexicon of individual speakers. Inter- and intra-individual variation likely plays a role across a range of factors, but little is known about the details.

- For a language learner to succeed in a wide range of communicative contexts, vocabulary in SLA needs to grow in two dimensions: Diversity (many words and productive combinations) and specialization (correct choices respecting semantic, idiomatic, and register constraints). Since those two are partially contradictory – if something is overrestricted, it cannot be sufficiently extended and vice versa – they are unlikely

to develop evenly throughout acquisition. This is further complicated since a learner also requires construction-specific knowledge. For example, a correct estimate of the degree of productivity of a given construction is required in order to avoid overextension not only in the lexical coselection, but at the very interface of lexicon and syntax.

- Item-specific knowledge can only develop through exposure to idiomatic input and over time, even more so if it is only partially taught in second language classrooms.
- A beginning learner in a structured acquisition setting is likely to be dealing with few grammatical structures and a small vocabulary. The combinatorial power of their vocabulary is low, and early acquisition stage material often works with tasks that require the learner to replace one or few words in a phrase, essentially offering full phrases for lexicalization. Learners at this stage therefore have few options to coselect outside of the expected.
- Learners at advanced stages in a school setting have accumulated experience that serves as a source of predicting the idiomatic choice in the target language through reading, listening, consciously learning collocations and looking up idiomatic translations of expressions from their L1. They may even have been exposed to a large amount of L1 target language input spending time in an immersive setting. Although it is known that near-native levels of idiomaticity are rarely reached, advanced learners still should have developed their idiomatic skill to a degree, i.e. limit their coselections to the more native-like ones to a larger extent from some point on, despite a formally higher combinatorial power.
- At intermediate stages however, learners are confronted with a large number of grammatical structures on different levels of abstraction, as well as a large number of new and more specific words, than those at early stages, and also the necessity to apply both to much more complex communicative contexts. The combinatorial power of their lexicon is much higher than at early stages. Assuming that there are no transparent semantic rules guiding nativelike selection, they cannot predict which items are likely to go together in a nativelike manner. This is equivalent to saying the combinatorial power of their lexicon is less constrained by other relations, and therefore realized to a larger extent. L1-transfer will also likely play a role especially in contexts where the learner's L1 might be particularly formulaic (loan translations, i.e. literal word-for-word translations of conventional coselections from the L1). The coselectional patterns of intermediate learners should therefore be most erratic.
- In effect, coselectional constraint should form a u-shaped development curve, or rather a u-shaped curve that is skewed towards a higher level at advanced stages vs. early ones, more resembling of a checkmark. U-shaped trajectories have been discussed for a number of phenomena from language acquisition and other learning processes (Plunkett and Marchman, 1991; Namy et al., 2004; Carlucci and Case, 2013). Carlucci and Case (2013) even suggest they are a necessary characteristic of the learning of systems that contain both general rules and idiosyncrasies. U-shaped trajectories are generally signs of processes of reorganization, where once established general rules require reformulation to allow for more specific hypotheses, thereby temporarily losing accuracy.

- It has been discussed in the literature whether verb-argument coselection should be modeled as an item-based, rule-based, or distribution-based phenomenon regarding the system in the *langue*, in usage, and in aquisition (Stefanowitsch, 2011; MacWhinney, 2014, among others). Distributional preferences in learning and generalizations in children and adults have been shown for novel verb and construction acquisition (Goldberg et al., 2004; Casenhiser and Goldberg, 2005; Wonnacott et al., 2008, 2017) and are a general feature of category acquisition and conceptualization (see section 2.2.2). The idiosyncratic nature of coselectional patterns and evidence from early FLA on the other hand seems to suggest an item-based path, as does the difficulty of defining distributions of coselected items for particularly rare items. Of course a dynamic or dialectical interaction of both is also plausible, albeit much harder to model. Potential cross-systematic rules have not found much attention since the phenomenon appears saliently idiosyncratic. However, some recent work suggests that fine-grained rules or rules on levels that are not generally taken into account, such as phonotactics, may play a role (Ambridge et al., 2012, 2014; Wonnacott et al., 2017).

- Assuming a process of lexicosyntactic and lexical diversification, both an item-based and a distributional account would start with few lexemes and end in many, but coselectional constraints would develop differently. In an item-based account, a core L1-like vocabulary in early acquisition would be expected. Items would then be broken up similarly to what has been described for early FLA, and rearranged in a nativelike fashion more or less evenly over time. In a comparison of text from different acquisition stages, this would be expressed in a long-lasting plateau of nativelike coselections through intermediate stages, but with a high rate of idiosyncratic combinations, and a more or less linear development of integration of more nativelike coselections into the lexicon. This is at odds with the assumption of a u-shaped development.

- A distributional account can be understood relative to linguistic categories or relative to interlanguage categories of a single learner or a group of learners. This makes it rather difficult to outline and trace categories precisely, because they may also differ between learners and language groups or by other factors. On a more general note however, a distributional account would be reflected in a more bursty development, where groups of coselections are broken up and recombined at a time, allowing for more combinatory freedom, and a sudden drop of nativelikeness. This is more consistent with the assumption of a u-shaped development. Distributional recombination should become visible in bursting increases of the rate of nativelike selection, but an overlap of several developments may hinder clear sight of this in the data. Measuring this would require genuinely longitudinal data and larger corpora, because effects of burstiness may be masked by inter-individual variation in a quasi-longitudinal or cross-sectional design.

- In either case, early L2 should be more homogeneous overall, followed by a phase of lexical diversification, where coselection at advanced stages is of a different quality. It should then affect not only a shared core vocabulary as in earlier stages, but also less frequent and more complex lexemes and coselections rather than being highly repetitive, but lexicosyntactically simple or trivial.

- Adult or nearly adult L1 speakers have had higher access to idiomatic input across

linguistic contexts and therefore develop a higher level of coselectional constraint than learners of any stage. Their language usage is also overall more streamlined and less sensitive to external factors such as L1- or L2-transfer in learners, cognitive fatigue, or contextual knowledge, resulting in less variance or higher homogeneity in L1 subcorpora vs. L2.

- In a semantically motivated approach, different argument slots are expected to show different degrees of flexibility. This is discussed in Plank (1984), suggesting that subjects are most flexibly exchangeable, while direct and indirect objects are less so, and prepositional objects are least flexible. Plank argues from a feature-semantic perspective: A verb such as *write* can take as an accusative object only nouns that extend to textual objects, like letters, books, messages, or music: ??*I wrote an apartment.*[1] However, adding a thematic relations perspective, all accusative objects share the property of 'being treated in a certain way' – written, read, bought, etc., thus anything that can be a semantic PATIENT can potentially occur in the slot of a relatively light verb like *to have, to give.*[2] Indirect or dative objects are further limited to a class of nouns that can act as a RECIPIENT, BENEFICIARY, or ADDRESSEE: ??*I gave the apartment the key.* However, this does not limit nouns simplistically: Consider the example *Dem Auto fehlt ein Reifen* ('the car is missing a tire'), where *Auto* ('car') would likely not be considered a potential RECIPIENT, BENEFICIARY, or ADDRESSEE in a context-free categorization. The argument still holds, though, because statistically, it would be expected that accusative object slots are less selective and require less metaphorization or creativity to be filled with many different nouns compared to dative slots, or prepositional objects.

  Prepositional objects in German are an interesting phenomenon. These are prepositional phrases that have object status but cannot always be clearly demarcated from free/non-obligatory adverbials, as in *sie läuft* 'she is walking, running' vs. *sie läuft auf die Straße* , 'she is walking, running into the street'. If a verb, however, governs specific prepositions and with it changes verb sense – as in *verstehen unter* ('to categorize, to view (literally: to understand under)', this is usally marked a prepositional object (for a comprehensive syntactic analysis, see Breindl (2011)). Thus, through the view of a specialized verb sense bound to the choice of preposition, a semantic specificity is expected from a prepositional object that should, with Plank, limit coselectional freedom. A special case in this regard are support verb constructions, *Funktionsverbgefüge*, such as *in Erwägung ziehen* ('to consider (literally: to pull into consideration)'). Many support verb constructions (but not all) in German contain a prepositional object, and support verb constructions are obviously coselectionally constrained. Conclusively, prepositional objects should be the most constrained class.[3]

[1] Other readings are of course possible: Resultatively: *I wrote a book and it has paid for my apartment*; or metaphorically: *I am writing a novel – last week, I wrote three main characters, today, I wrote the apartment.*

[2] For more in-depth treatment of semantic or proto roles, see Stevenson et al. (1994); Baker (1997); McRae et al. (1997).

[3] Genitive objects have become very rare in modern German and are intrinsically coselectionally constrained through their rarity (what occurs only once or twice cannot be many different nouns). Theoretically, the most obvious role of the genitive object in German would be THEME (*Wir gedenken seiner*, 'we commemorate him', *Sie beschuldigte mich des Verrats*, 'she accused me of betrayal'). Since THEME is a very broad label, thematic role constraints are unlikely for those. However, since many previously genitive objects have turned into dative objects in present-day German, the verbs governing

This argument is rather structural and belongs to a syntactico-semantic model. Zeldes (2012) shows varying degrees of productivity for different argument slots empirically. He adds, however, as it was also discussed in the previous chapter, that syntax does not fall neatly into purely semantically driven categories. Rather, idiosyncrasies or not yet described rules play a role in the case of verbs in verb-argument distributions and in the productivity of verb-argument coselection. It is therefore likely that object type slots and other categories, like verb types or lexical clusters, compete in determining the coselectional constraint of a specific verb. In any case, since much linguistic theory predicts different behavior by argument slot, a distinction should be made in the analysis, where possible.

Given the limited amount of formal modeling and operationalization of related concepts that has been done so far, especially in terms of distributional vs. item-based accounts in corpus research and learning trajectories, this study will be limited to some core aspects of the discussed. Under the umbrella of the wider research question of "(How) can the development of lexicosyntactic constraint be shown as a structural property of L2?", the hypotheses guiding the corpus study are therefore as follows:

1. A process of lexicosyntactic diversification from early to late acquisition stages is visible in the data.[4]

2. Lexicosyntactic constraint as expressed in the variety of combinations of verbs and their argument lexemes is situated on a trajectory from beginning learners to intermediate to advanced.

3. Constraint is lower at intermediate stages than early and late stages in L2, such that advanced learners are most and intermediate learners least similar to L1.

4. Coselectional constraint is lower in L2 than L1 at all acquisition stages.[5]

5. Subcorpora in L2 have higher intra-group variance than L1, and variance is lower in subcorpora of beginning learners vs. intermediate and advanced ones.

Coselection here is understood as a lexicosyntactic rather than a positional phenomenon, which means that the analysis is based on the coselection of verbs and their argument slot lexemes. For example the verb *hören* ('listen') and its accusative object *Musik* ('music') are considered a coselectional pair or simply coselection.

There are linguistic arguments in favor of a positional model in which coselections are a category of adjacent or nearby words, such as a more traditional notion of collocation

the remaining ones are often slightly antiquated, like *sich entledigen, sich rühmen, sich bemächtigen*, 'to dispose of', 'to pride oneself in', 'to take possession of'. These should then be limited to specific registers and with it nouns likely to occur in those.

[4]This may seem trivial and even impossible to avoid, but it will become relevant in the analysis and the choice of method and is thus mentioned here.

[5]It could be argued that coselectional constraint in L2 is equally strong as in L1, but with different lexemes (overuse of frequent collocations, chunks, *lexical teddy bears*; see chapter 2.2). At the same time, with a higher degree of specialization and or through larger vocabularies, native speakers should still show higher coselectional constraint in total or on average *despite* also being more productive when needed. In other words, learners are presumed to recombine meanings of which they cannot access the conventional form, and tend to recombine meanings from their limited vocabulary. Native speakers can both access more conventional or acceptable forms and can also use larger vocabularies when productivity is required, thus producing more specialized new forms.

would suggest. Some of these arguments are connected to concepts like entrenchment, entailment, or the holistic storage of phonetic or auditory chunks. These are reflective of a process-oriented prespective and from a lexicosyntactic perspective mostly relevant within a lexicalist framework, referring to grammars such as pattern grammar, where syntax is essentially modeled as an epiphenomenon of the lexicon (see Hoey (2012); Hunston (2012); Hudson and Hudson (2007); and Müller and Wechsler (2014) for a critical analysis).

On the other hand, the literature review in the previous chapter has shown coselectional phenomena to affect abstract levels of language to a much larger degree than was expected at first in the study of phraseology, and a lexicalist account poses intrinsic problems to a linguistic categorization of which lexical items and which argument structures may be compared. I will therefore use categories of verbs and their NP and CP arguments. Which arguments are considered in each analysis and why will be discussed separately where necessary. Eventually, a synthesis of the more phrasal and the more lexicalist approaches might prove most fruitful for the understanding of coselectional constraint, but this requires intricate modeling on several linguistic layers at once and cannot be done within the scope of this work.

Further interesting hypotheses can be derived from the theoretical background, such as how the semantic specificity and productivity of a verb in L1 interact with the phenomenon in L1 and L2, and what to expect from typological aspects of the L1 of participants, some of which will be discussed in chapter 7. This study will however be limited to showing whether and how lexicosyntactic constraint as a structural property of native and learner language can be measured, and how this can be done in small to medium-sized corpora specifically.

3.2. Data

The data used in this study has been collected by the Kobalt research network (Zinsmeister et al., 2012) and contains a total of 151 essays written by Belarusian and Chinese learners of German (henceforth BEL and CH), and 20 essays written by native speakers of German (high school students from Berlin, 12th grade, *Grundkurs*[6]), henceforth L1. Individual documents are referred to with either BEL or BY for Belarusian learners and either CH or CMN for Chinese learners, and DEU for L1, and an identifying number, such as CMN_017, BEL_020, or DEU_010. Documents can be identified in the publically available corpus in the same way.

All L2 participants were university students majoring in German at their respective universities in Belarus and China, and there is extensive metadata indicating first and second languages, acquisition time and experience in L2-speaking countries for German and other languages. Essays were handwritten in reply to the prompt *Geht es der Jugend heute besser als früheren Generationen?* ('Are adolescents today better off than previous generations?') under equal conditions (90 minutes, no aids like dictionaries allowed).

All participants from the L1 and L2 groups took the standardized cloze test onDaF (now onSET) that is part of the TestDaF battery, an official German skill certification program comparable to the English TOEFL and IELTS, Spanish DELE, and French DELF/DALF.

[6] *Grundkurs* refers to a class level that is chosen by students in German high schools as opposed to *Leistungskurs*. Most subjects are taught as *Grundkurs* in three weekly lessons, grade count less towards the final average grade, and final exams are usually facultative, whereas two to three self-elected *Leistungskurse* are taught in five weekly lessons, have a higher impact on the final grade, and final exams in the chosen subjects are mandatory.

Specified TestDaF levels representing a language skill level from the Common European Framework for Languages (A1–C2, henceforth CEFR, Council of Europe (2017)) are required to study at most German-speaking universities and some German study programs outside of German-speaking countries.

The original aim of the Kobalt project was to compile a small, but deeply annotated and strictly homogeneous corpus in terms of learner backgrounds and level of German. This is why in the original dataset that was processed by members of the project, only 20 texts per language were considered, all within a range of 114–129 onDaF points (roughly equivalent to the upper intermediate B2.1 in levels of the CEFR.[7]

Beyond the 20 texts within the onDaF range in each language group, the Kobalt team has collected 111 texts written by learners from China and Belarus who scored above or below the onDaF score limit. Together with the base corpus, those make a total of 171 texts (87 BEL, 62 CH, 20 L1). These are distributed across a range of 34 to 148 onDaF points in BEL, 72 to 148 points in CH, and, relevantly, 133 to 154 points in L1.[8]

The original Kobalt data also included 11 texts written by German learners with L1 Swedish, but no texts beyond original onDaF limit, which were therefore excluded from the analysis here.

3.2.1. onDaF-based grouping

This study is cross-sectional divided by a scalar variable that is presumed to correlate with proficiency. This design in the case of SLA is sometimes also referred to as *quasi-longitudinal*, as opposed to truly longitudinal data that documents the acquisition of a cohort of speakers. The advantage of a cross-sectional or quasi-longitudinal design is obvious: Data from several levels of proficiency can be collected at once, there is no necessity to wait and hope for low dropout rates and high success rates in learners, and individual quirks are less prominent in the total dataset. On the downside, the argument lacks a *truly* developmental perspective, since changes cannot be observed, only results of presumed changes. In other words, learner A may use a structure half as many times at score level s_1 than learner B at score level s_2, but this cannot count as proof that the structure is learned to be used more often between those two score levels. At the same time, there exists the idea in the *interlanguage* hypothesis (Selinker, 1972) that language learning happens in a continuum from beginner to advanced to, in rare cases, near native-like, and that systematic changes happen on this trajectory; and that these systematic changes are *not* nothing more than clustered idiosyncrasies of individual learners, but reflect structural processes in an interlanguage space. That is the perspective taken in this thesis.

Thus, to determine the order of learners on said trajectory and as a grouping variable, onDaF test scores are used.[9] This is not without problems, which I will briefly discuss in

[7] See Council of Europe (2001) and Council of Europe (2017) for a description of items and an identification of the purpose, hierarchy and corresponence to linguistic concepts of the levels

[8] This means that none of the native speakers reached virtually perfect scores as would be expected from a valid c-test, and that some of the native speakers scored barely higher than is expected by higher-intermediate ranging learners and much lower (14 points) than learners from both L2 groups. This suggests the test was either particularly difficult or particularly confusing. The validity of the instrument for this study will be discussed further below and in chapter 6.

[9] It may also be possible to cluster learners by syntactic or other features, thus simulating more complex language assessment. This would make for a circular argument if lexicosyntax was involved, though, because if clusters are determined by lexicosyntactic features, they will also exhibit those in the analysis; and, as will be argued in this section, acquisition stages are a problematic concept in terms of the

this section. I will conclude that the approach is valid, as long as onDaF scores are not used as a unique identifier for language acquisition stages, but for ordering and grouping.

The onDaF test, short for *online TestDaF*,[10] is a standardized cloze-test (c-test) consisting of eight short texts of which the second half of every other word is deleted. Participants can reach a score between 0 and 160, where each point stands for a correctly filled gap, and texts are progressively more difficult in terms of syntax, lexicon, and register. While much can be critized about the use of cloze-tests for the assessment and testing of general language skills, the onDaF has in fact been validated against the skill-specific tasks of TestDaF. Eckes and Grotjahn (2006) and Eckes (2017) report high correlation and a one-dimensional construct in validation, which means that the construct measures one dimension and not several (extrapolating this dimension is 'general language skill').

The main problem with this is that 'general language skill' is in itself not a very clear construct, and there has been an ongoing debate about the correlation between test scores, assigned CEFR levels and language skills as described by CEFR. As Wisniewski (2017a,b) points out, learner language can rarely be described as belonging to a single level on all dimensions of the test (writing, reading, speaking, listening), especially when it comes to the formal linguistic correlates like the use of specified syntactic or lexical material. Rather, while scoring agreement is high, and scoring constructs are well-defined, the constructs that are *actually* scored by raters seem much less clear and objective, and differ even when raters agree on a score. This has been noted and discussed for a long time, partially because it is up for question whether communicative skill is different from other constructs typically measured in psychometrics. Language in general shows a high degree of variation in its highly and complexly interwoven syntactic, morphosyntactic, lexical, textual, and pragmatic subsystems. Different, even complementary skills may level out weaknesses in a way that makes it *objectively* hard to grasp what makes for a learner's actual or perceived high or low communicative competence (Swain, 1993). In fact, the high statistical validity and the empirical acceptance of cloze-tests for placement tests in general shows that it does not always matter *which* gaps a learner fills correctly, but that scoring within an approximate window yields information about 'how much' of the target language they are able to process.

Of course in language assessment and SLA research alike, many more critical questions have been raised: Whether describing language acquisition in terms of stages is valid in itself or whether language development is too much of a discontinuous phenomenon on the different linguistic levels (Young, 1995; Perkins et al., 1996); whether, rather than assigining a combined label to a group of skills, localization on individual skill-based interval scales would be more accurate; whether learner varieties are simply too multi-dimensional to be grouped into level labels (see also the broader discussion of the notions *variety* and *interlanguage*, Dimroth (2012); Han and Tarone (2014)).

With this, the onDaF may not give precise information about the general and specific skills of learners. And it may be a somewhat imprecise instrument when it comes to the details. But it should also be considered that not all linguistic questions that require an ordering of learners into higher and lower skill groups also require an exact assessment of their skills on all dimensions; and that the onDaF is still a validated language assessment that provides orientation regarding the location of a learner in an interlanguage space, which is rather difficult to come by with limited resources. Perhaps the difference between

demarcation of clearly defined linguistic correlates.

[10] The test has since been renamed to onSET for *Online-Spracheinstufungstest* ('Online language placement test'), but for the purposes of this work I will use the name used at the time of data collection.

113 and 117 onDaF points is not very revealing of skill differences between two learners, but a difference between 110 and 130 certainly is, and especially so since the eight cloze texts in onDaF are progressively difficult to solve, which means that overscoring by chance is progressively unlikely. This is further supported by the fact the onDaF instance used in the data collection here must be a particularly challenging one, which is reflected in the low-scoring native speakers. It might on the other hand also be a particularly confusing one, which makes it hard to assess how well it distinguishes between the more advanced learner stages or groups. This cannot be resolved with the data here, and in a pragmatic compromise between resource efficiency and expected reliability, the onDaF will be used as a ranking and grouping variable despite these shortcomings. This will first be in groups of fixed onDaF ranges: below 75, 75-94, 95-114, 115-129, 130 and above; and later as ranks (lowest in a subcorpus = $rank_1$, highest = $rank_{subcorpus\ size}$ (see chapter 6.3.2.2).

The onDaF does not scale linearly, so that a range of 20 points is not necessarily 1/3 larger than a 15-point-range in terms of language development – it is in fact quite unclear what a linear development of target language profiency could denote (what is '1/3 more target language'?). Neither do these score ranges correspond ideally to CEFR-levels as reported and used by the TestDaF publisher. Instead, they were chosen pragmatically to work around the original Kobalt corpus and to maintain sufficiently large groups in the extended corpus. Ranges still correspond roughly to CEFR-levels as they are reported for the English equivalent (A2, B1, B2.1, B2.2, C1 and higher), but these may deviate from German thresholds significantly. It was unfortunately impossible to transparently report score ranges relative to German CEFR-levels as used in the official test due to publisher restrictions. However, with a lack of a model of acquisition stages in the first place, and the uncertainty regarding the role of CEFR levels and linguistic correlates, this is not of particular interest. No claims are made regarding coselectional constraint "at B1-level", for example. Instead, the scale is used as an interval scale for grouping and comparison.

In summary, the groups, while they do represent progress in language learning, are not theoretically mapped to any specific theory of acquisition stages or combination of skills, but rather a necessary grouping of a continuous variable into an interval scale for comparability between corpora of certain size. I will discuss aspects of the validity and ways around this approach at length in chapter 6.[11] Tables 3.1 and 3.2 give an overview of the number of documents and tokens in the Kobalt subcorpora resulting from the described onDaF grouping.[12] All texts considered, Kobalt is a mid-sized corpus with overall 105 668 tokens and 93 179 tokens in the L2 part (tokens are counted on ZH1, see next section). If texts are divided into onDaF groups however, each group results in a small subcorpus of less than 30 000 tokens for each onDaF group for both L2 groups combined, and an even smaller subcorpus of under 15 000 tokens for most language- and onDaF-group split subcorpora. The limitations of this data size will be discussed in the chapters 6.3.2.2 and 7.2.3. However, while the chance of reaching convergence of category probabilities are better in large data, smaller data provides a better overview of the text before abstraction, since it can be fully read; and more control over the analysis. More will be said about the advantages and disadvantages of small to medium-sized corpora in chapter 7.2.3.

As can be seen in the historgam in fig. 3.1, setting different onDaF ranges for the grouping would not have yielded more balanced group sizes because scores are clustered

[11] The analysis will show that group effects are in fact larger than individual effects, and that a grouping by onDaF ranges is not inferior to a purely scalar approach, i.e. a sliding-window-sampling of each 10 (15; 20) texts in ascending onDaF order, see sections 6.3.2.1–6.3.2.2.

[12] OnDaF group 130 is the original Kobalt corpus except for three texts that reach 114 onDaF points and were reassigned to the 115 onDaF group here.

language \ onDaF group	75	95	115	130	160	sum
BEL	11	27	21	20	10	89
CH	–	10	24	17	11	62
L1	–	–	–	–	20	20
sum	11	37	45	37	41	171

Table 3.1.: Number of documents in Kobalt subcorpora. Numbers in onDaF group refer to the upper score limit, i.e. 75 = onDaF score < 75, 95 = 75 ≤ onDaF score < 95, etc.

language \ onDaF group	75	95	115	130	160	sum
BEL	3 328	16 210	14 513	14 609	8062	56 722
CH	–	5 542	14 062	10 300	6 553	36 457
L1	–	–	–	–	12 489	12 489
sum	3 328	21 752	28 575	24 909	27 104	105 668

Table 3.2.: Number of tokens in Kobalt subcorpora

around roughly 95, 115, and 135 points in both languages and not distributed evenly elsewhere. As far as text length is concerned, fig. 3.2 shows that texts grow increasingly longer with higher onDaF in BEL, but not in CH. I will discuss repercussions on the linguistic model, approaches to a text length normalization, and appraoches to a validation of results from varying text length in chapter 6.3.4.

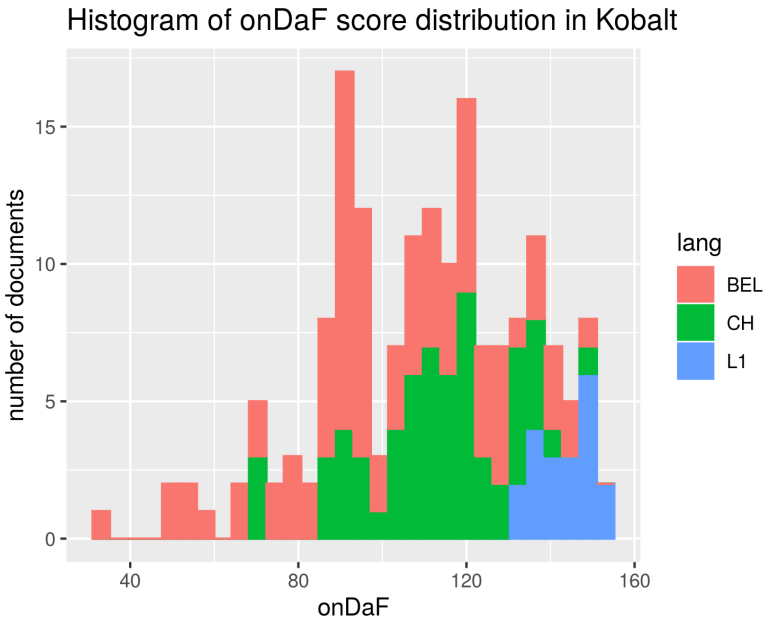


Figure 3.1.: Histogram of the onDaF score distribution

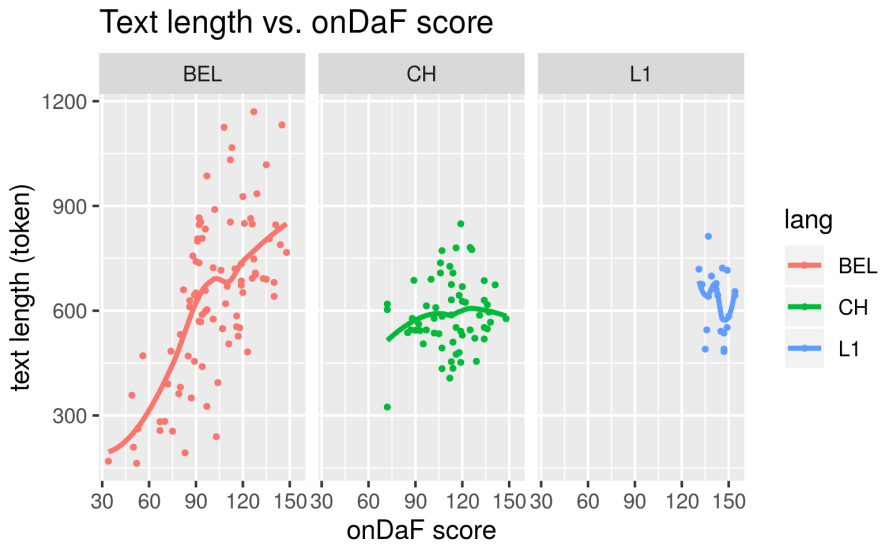


Figure 3.2.: Text length distribution vs. onDaF scores

3.2.2. Annotations

Data from the original Kobalt project was transcribed independently by two transcribers and versions compared and corrected by a third one. It was then part-of-speech-tagged (POS-tagged) and lemmatized with TreeTagger (Schmid, 1995) and dependency-parsed using an instance of the Maltparser (Nivre et al., 2006) trained by the corpus linguistics working group at the Humboldt University of Berlin. Parses were made on target hypotheses 1 (target hypothesis tiers are labeled *ZH0/1/2* for *Zielhypothesen 0/1/2*) and manually corrected.

Target hypotheses were first developed for the German learner corpus Falko (*Fehlerannotiertes Lernerkorpus*, (Reznicek et al., 2010, 2013)) as an orthographic and syntactic normalization layer that improves automatic processing and findability of phenomena (Lüdeling, 2008; Lüdeling et al., 2005). Target hypotheses 1 are strictly regulated in annotation guidelines that are documented in Reznicek et al. (2010), whereby no semantic material is changed or added, and missing subjects or objects are filled with functional objects like *es* or *das* ('it', 'this/that'), grammatical word order and morphosyntactic congruence in case, number, and tense, aspect, mode (henceforth TAM) are established, and orthography is corrected. The advantage of target hypotheses over error correction in the original text lies in the visibility of competing target hypotheses and the explicitation of assumptions about the target structure.[13] As has been mentioned, no semantic material is changed or added in ZH1, which is relevant to the study at hand. This means that odd choices or coselections that make little sense semantically are part of the data and were not excluded from the analysis.

In the study here, only target hypotheses 1 are used, but the original Kobalt data also

[13]It is impossible to tell definitively what a learner meant or set out to express. Even ungrammatical structures that appear quite obvious such as disagreement in number or case in inflections may be resolved by either changing the noun or the verb, and it is helpful for quantitative analysis to skew in one direction while also keeping changes transparent. An extra corpus tier is added to automatically document changes from the original text to the target hypotheses using tags such as DEL (deletions), INS (insertions), and MOVS/MOVT (source/target of movement).

includes *ZH2 – Zielhypothese 2* (target hypothesis 2), referring to a stylistically and semantically corrected, native-like version of the text, and *ZH0 – Zielhypothese 0* (target hypothesis 0), a version of target hypotheses 1 without constituent movement. Furthermore, topological field and discourse annotations as well as a wide range of metadata including a number of NLP measures for complexity and other features are available. Recently, rhetorical structure analyses based on RST (rhetorical structure theory) have been added as another annotation layer (Wan, in prep.). The original Kobalt corpus is available at `http://korpling.german.hu-berlin.de/annis3/#_c=a29iYWx0TDJ2MS40` (L2) and `https://korpling.german.hu-berlin.de/annis3/#_c=a29iYWx0TDJ2MS40` (L1) and can be further extended and enriched with new annotation layers at any point. Processed data including all scripts will be made available through `www.zenodo.org`, a repository dedicated to open science that offers long-term storage for research data (`10.5281/zenodo.3584091`).

For the additional data, due to lack of resources, essays were transcribed only once and only target hypotheses 1 were added based on the same annotation guidelines (Reznicek et al., 2010). Those were tested exhaustively in several learner corpus projects and do not leave much room for interpretation, but minor mistakes cannot be ruled out.[14] The data was also POS-tagged and dependency-parsed in the same manner as the base data.

For all data including the original Kobalt corpora I manually corrected dependency parses and POS tags in the verb-argument complex in the following ways:

- TreeTagger POS-tags in the verb domain, such that VA/VV/VM and FIN/INF/ PP were correctly assigned. TreeTagger tags accurately between larger part-of-speech categories such as verbs vs. nouns vs. prepositions and so on, but has difficulty choosing between the finer distinctions in learner data. This particularly affects the infinitive vs. finite verb distinction, but also the distinction between lexical and auxiliary/copula uses of *haben* ('to have'), *werden* ('to become'), *sein* ('to be').

- TreeTagger lemma tags for unknown lexemes, mostly referring to technology or newer developments in society that the TreeTagger dictionary from the mid-90s does not cover, and polysemous verb forms as in

  (1) Nun soll abgewogen werden, welchen Wert [diese Probleme] hatten
  Now shall evaluated be.pass which value [those problems] had
  'Now the value of those problems shall be evaluated', *DEU_020*

  where *abgewogen* is tagged as *abwägen/abwiegen* ('evaluate, estimate | weigh') and disambiguated to *abwägen* ('evaluate, estimate') in the data, since unlike its English counterpart, German *abwiegen* and *abwägen* cannot be used synonymously.

- Dependency labels in the verb-argument complex only (verbs, objects, subjects, prepositional phrases vs. prepositional objects. Generally, wherever dependency labels were changed, this was done with the goal of keeping interesting cases findable.

[14] I believe I can rule out major mistakes because I checked annotations carefully several times during and after transcription and annotation, and later, while correcting dependency tags and adding verb category annotations, I re-evaluated some of the transcriptions and annotations for correctness and plausibility, and performed several sanity checks in the process of data processing and analysis. Still, it would be preferable for other researchers willing to use the data to double-check and eliminate mistakes that may still exist.

This is not to claim an ideal syntactic analysis, which would be an ambitious endeavor using only the framework of dependency grammar in the field of lexicosyntax specifically. Instead, differences in lexicosyntactic analysis are *marked out* through annotations, without being *ideally labeled or categorized*. This affects primarily the following aspects: Generous assignment of verb status to present participles (see next bullet points), predicate status to presumed obligatory constituents in constructions (*besser*, 'better' in *es geht ihnen besser* 'they are doing better', a construction of the verb *gehen*, 'to go'), and prepositional objects (*OBJP*) whenever a PP appeared obligatory or replaced an obligatory predicate or adverbial phrase. Consider the following example from Kobalt:

(2) Einige davon sind eingebildet und sozial ohne Pflichtbewusstsein.
Some of_them are arrogant and socially without responsibility
'Some of them are arrogant and show no sense of social responsibility' (*CMN_017*)

Here, *eingebildet* 'arrogant' is a simple adjectival predicate, but *ohne Pflichtbewusstsein* 'without responsibility' is a prepositional phrase, which cannot theoretically be construed a prepositional object (since the copula *sein* 'to be' does not take objects (outside of specific constructions), but only links predicates). It should be a complex or prepositional predicate. But assigning it a PP makes it impossible to distinguish it from other PPs which are adjuncts but still dependents of the verb, such as *Sie fahren [mit ihm] nach Hamburg* 'They go to Hamburg [with him]'. To keep these distinguishable, I marked them OBJP. There are some other syntactic cases that were treated in the same way, like constructions and mandatory local adverbials or directional objects as in the following example.

(3) Aber die Jugendlichen (...) setzen sich immer in das Zentrum
But the youth (...) sit themselves always in the center
'But adolescents (...) always put themselves at the center' (*CMN_017*)

Obligatory adjectival and adverbial complements to construction-marked verbs were labelled as PRED.

- Assignment of dependents to heads, where all dependent words of the same kind were assigned the same label, such that a verb can have several OBJA or OBJP. This is divergent from the dependency grammar developed by Foth (2006), where argument slots can be filled only once and all further listings are assigned coordinated complement labels such as CJ and KON. I changed this for easier subsequent processing, but original parses are preserved and a corrected version in line with (Foth, 2006) can be reverse-engineered, if necessary. With similar intent, subjects were assigned as dependents of the lexical, not the finite verb in auxiliary and modal constructions. For more details and examples see chapter 5.

- PPs were assigned to the verb in ambiguous cases whenever they could plausibly be placed in the *Vorfeld* (first constituent before the finite verb in German main clauses, see Ramers (2006); Reis (1980)). It should be noted that no concrete error estimation has been performed, but, particularly in the low-scoring onDaF range, parser output was rather inaccurate with an estimated error rate of at least 30%. In all cases, subjects were assigned to finite verbs and objects and other dependents to lexical

verbs in complex verb constructions, and lexical verbs as dependents to auxiliary and modal verbs. Since each word can only have one head in dependency grammar, with coordination of verbs, some end up subjectless in the analysis although the subject is syntactically and semantically specified:

(4) Sogar heute in dem 21. Jahrhundert verhungern die Leute, leben in
Even today in the 21st century starve the people, live in
schrecklichen Bedingungen und müssen damit allein zurechtkommen
horrible conditions and must with_that alone cope
'Even today in the 21st century, people starve, live in horrible conditions and must cope with that all alone' (*BEL_020*)

In this sentence, *Leute* ('people') is only assigned dependency to *verhungern* ('to starve'), although it is also the subject to *leben* ('to live') and *zurechtkommen* ('to cope'), see fig. 3.3.

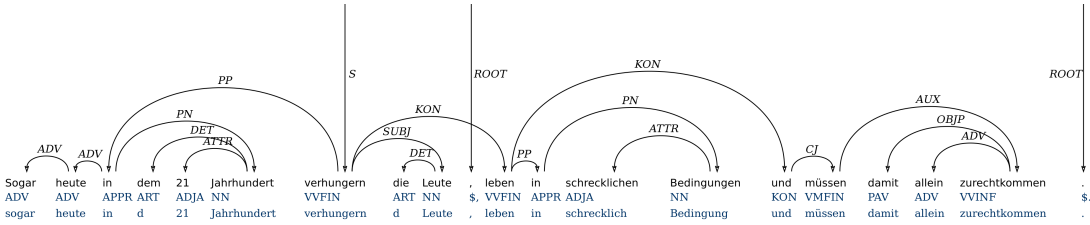


Figure 3.3.: Dependency parse of a sentence with a chain of coordinated verbs. The subject is only assigned to the first verb since each token can only depend on one other token (*verhungern*, 'to starve'). All dependency parse visualizations are done with `http://arborator.ilpga.fr/q.cgi`..

- Participles are a problematic case in German. Located at the interface of morphology, syntax, and lexicon, they can be interpreted as verbs in state passive, as deverbal adjectival predicates, or entirely lexicalized adjectives (Lenz (1993); Fuhrhop (2012, chapter 8)). Consider ex. (2) above, where *eingebildet* 'arrogant' is derived from *einbilden* 'imagine' and homonymous to 'imagined' as in 'they had only imagined that' (*Sie haben sich das nur eingebildet*). Sometimes such participles open argument structures as defined by the signature of the underlying verb:

(5) Die Frauen oder auch Mädchen wurden als den Männern und Jungs
The women or also girls were as the.dat men and boys
untergeordnet betrachtet
subordinate.participle view.participle
'Women and girls were seen as inferior/subordinate to men and boys' (*DEU_012*)

In this example, the subordinate phrase *den Männern und Jungs untergeordnet* is a complement to *als*, and *den Männern und Jungs* is a dative object to the verb *unterordnen* ('subjugate, subordinate'). However, like in the previous example with *eingebildet*, *untergeordnet* can also be used as an adjective without the underlying verb's argument structure. In fact, for many adjectives the deverbal reading is not the most frequent or the first that comes to mind (as in the case of *eingebildet*). I

have therefore decided to treat deverbal units as passives when they govern phrases based on the argument structure as part of the verb signature, and treat them as adjectives in other cases. This means, that in my model, there are two very closely related categories: One includes copula verbs and many different predicates, including some participles. And the other includes cases where those participles also appear as state passives with the same verb (*sein* – 'to be'), but are analyzed not as a copula, but as auxiliary verbs, once the participles show their own argument structure. This can be viewed as a minimizing strategy on the verb side, since verb lexemes are only counted when they appear with arguments. It also avoids an overestimation the complexity of the verbal phrase through excessive passives in phrases that are parallel to simple nominal or adjectival predicates. It is not, of course, an ideal model of the German participle, but a pragmatic choice to maintain access to identical coselections whether they occur in a participle-based structure or outside of one.

Additionally, verbs were annotated by nine morphosyntactic categories (copula, auxiliary, modal, modifying, simple lexical, particle,[15] prefix,[16] construction (cx: reflexive constructions, modal infinitives, and some specific uses of verbs like *gehen um* ('to be about')), and gehen_cx (*es geht der Jugend gut/schlecht* ('adolescents are doing well/badly'), treated separately from other constructions due to high frequency because it is part of the prompt, while also keeping it distinguishable from other uses of *gehen* ('to go')). Annotation guidelines for those categories as well as ambiguous or idiosyncratic cases of the corrections above can be found in the repository (`10.5281/zenodo.3584091`).

The full (extended) Kobalt corpus can be found in the repository and shortly will be made available through ANNIS (Krause, 2019), too. In summary, it contains contains partially corrected POS and lemma tags and dependency parses, where all corrections focus on verb-argument structures (VAS), and morphosyntactic verb category annotation, all based on target hypotheses as a normalization layer. Uncorrected tagger and parser output can also be accessed in separate tiers. See fig. 3.4 for the parsing visualization. The data was then parsed into a list of mothers and daughters (heads and dependents) using R on RStudio (R Core Team, 2015; RStudio Team, 2015) and its packages dplyr and reshape2 (Wickham et al., 2018; Wickham, 2007), so that the prompt *Geht es der Jugend heute besser als früheren Generationen?* ('Are adolescents today doing better than previous generations?') looks as illustrated in tab. 3.3.

[15] Particle verbs are complex German verbs. They are compounded from a base verb and a free morpheme like a preposition: *auf + schreiben -> aufschreiben* ('up' + 'write' -> 'write down, note'). Unlike in English phrasal verbs, the particle is incorporated in the verb in infinitive and participle forms, but splits from the verb in the finite form: *Ich habe das aufgeschrieben* ('I have written that down'), *Ich muss das aufschreiben* ('I have to write that down'), but *Ich schreibe das auf* ('I am writing that down').

[16] Prefix-derivations of verbs, like *setzen -> zersetzen, versetzen, besetzen*, etc. ('to put, to place, to sit' -> 'to decompose', 'to transfer', 'to occupy'. Unlike particles, prefixes are bound morphemes and do not move syntactically.

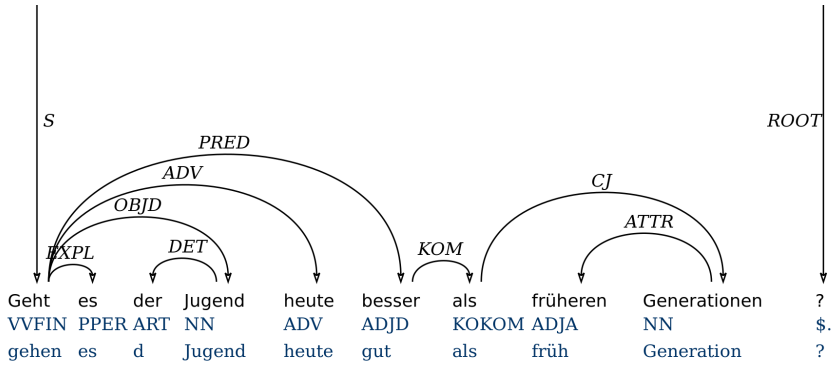


Figure 3.4.: Dependency parse of the prompt *Geht es der Jugend heute besser als früheren Generationen*

mother_lemma	mother_dep	mother_cat	daughter_lemma	daughter_dep
gehen	S	gehen_cx	heute	ADV
gehen	S	gehen_cx	es	EXPL
gehen	S	gehen_cx	Jugend	OBJD
gehen	S	gehen_cx	gut	PRED
Jugend	OBJD	NA	der	DET
besser	PRED	NA	als	KOM
als	KOM	NA	Generation	CJ
Generation	CJ	NA	früh	ATTR

Table 3.3.: The prompt *Geht es der Jugend heute besser als früheren Generationen* parsed into heads and dependents (mothers and daughters).

3.3. Summary

In this chapter, hypotheses were derived from the previously discussed theoretical background. The small to mid-sized German learner corpus Kobalt was presented and described in terms of the conditions of data collection and annotation choices, and the plausibility of the use of onDaF as a grouping and ranking variable was discussed. Further details regarding annotation choices will be discussed in chapter 5.

4. Statistics

In this chapter, the measurement of coselectional constraint as a structural property of language in native speakers and learners is approached statistically. With the aim of developing an operationalization of the research questions stated in the previous chapter, it seeks to answer the following questions:

- Can a process of diversification be observed?
- Can a process of specialization be observed?
- What are expressions of both in the data?

While it may seem trivial to expect both a diversification and a specialization to take place, it is not yet clear how these would be expressed. Different expressions may have different repercussions on the study of coselectional constraint; and, as will be shown, they do.

- How similar are texts regarding lexical and lexicosyntactic choices?
  - How are verb categories and argument types distributed in the various subcorpora?
  - How much of the vocabulary is shared between texts?

Coselectional constraint, as it is conceptualized in the *idiom principle* (Sinclair, 1991) and in phraseology, implies the identity of coselected items. If learners at different stages of acquisition all use different words, their coselections will also differ. Similarly, if learners all prefer different VAS and argument slots coselect differently, their coselectional constraint may differ as a result. This would require a different intepretation of the results compared to different levels of coselectional constraint within the same argument slot. Thus, an estimate of the lexicosyntactic similarity between texts is necessary for the interpretation of results.

Finally, the central research question is:

- How constrained are subcorpora regarding lexical and lexicosyntactic choices?

It will be shown that while the other questions can be answered from descriptive and inferential statistics (lexical association as expressed in ΔP, Gries (2013)), this final question cannot be usefully operationalized in a statistical approach in this data. This may partially be due to size. But primarily, it is because a statistical analysis relies on factor combinations, i.e. individual and concrete coselections. The two processes of diversification and specialization, however, play out in a way that makes it impossible to track the same items over time. Besides, the combinatorial power of coselections is huge despite large overlap in shared vocabulary between texts. Randomness cannot reasonably be used as a baseline (Kilgarriff, 2005) and no other quantifications of the lexicosyntactic constraint *in total* exist at present. It is thus unclear how to define high or low coselectional constraint against this background.

4.1. Diversification

Several aspects underly development in SLA:

- lexical and coselectional diversity;
- morphosyntactic diversity and complexity of TAM forms including modals, modal infinitives, and passives;
- syntactic development: diversity and complexity of verb argument structures (VAS), i.e. frequency of argument slot types; use of constructions.

Since coselectional constraint implies a certain 'sameness' of lexicosyntactic items between cohorts, these processes are relevant for the assessment of coselectional constraint and will thus be reported in this section.

4.1.1. Lexical diversity

Lexical diversity can be measured in a number of ways. To first give an impression of how many unique lexemes occur for each argument slot and verbs that occur with such arguments, see tab. 4.1 and 4.2. The tables show that both native speakers and learners use a large number of unique verbs, particularly with SUBJ and OBJA slots. They also show that native speakers use more unique verbs in OBJP and OBJD slots relative to SUBJ and OBJA slots compared to the learners; and that native speakers use many more unique SUBJ argument lexemes compared to OBJA, while learners of both groups use more unique OBJA lexemes compared to SUBJ.

Since these are not normalized by text length or corpus size, the absolute numbers cannot be compared. They mainly serve as an illustration of the lexical diversity at all stages of acquisition, even the low-intermediate ones: Despite consisting of only 10 texts, the CH-95 corpus contains a total of 84 verbs that occur with subjects; and, similarly, in the BEL-75 corpus, learners that would be placed in an A2.2 or B1 class if onDaF scores were used for placement, and also write rather short texts, come up with a total of 47 unique verbs taking an accusative object.

The highest degree of diversification is observable in the OBJP slot: In BEL-75, 11 learners use a total of only 13 verbs with OBJP slots. In BEL-160, 10 learners use 62 different verbs; as many as 27 learners in BEL-95.

Since text length varies considerably, a normalization is in order; and an estimate of the lexical diversity of individual learners, and variance between them is necessary as well. The most common and simplest metric that considers diversity relative to text length is type-token-ratio (TTR), where all unique types, in this case lexemes, are divided by the number of tokens in a text. TTR necessarily interacts strongly with text length due to the Zipf-distribution[1] of lexemes and the reuse of words in a topic once they are introduced. It cannot stagnate, but de- or increases for each new token depending on whether it has previously occurred or not. Since text length also correlates with onDaF in BEL, figs. 4.1

[1] It was mentioned in chapter 2.2.2 that the Zipf-distribution of lexical material in corpora has been called into question lately, see Piantadosi (2014); Aitchison et al. (2016). These papers do not dispute the observation that lexemes are distributed by some power-law function in corpora, but raise attention to the possibility that there are not yet well-understood underlying processes, latent variables, which result in the distribution. For the purposes of this chapter, the term will still be used as a descriptor of a power-law-like distribution of lexical items.

dep	CH-95	CH-115	CH-130	CH-160	L1	BEL-75	BEL-95	BEL-115	BEL-130	BEL-160
OBJA	74	142	138	97	148	47	155	148	168	140
OBJD	11	19	24	15	31	6	14	14	17	21
OBJP	35	73	69	50	102	13	62	62	85	62
PRED	3	6	5	3	5	3	4	6	4	3
SUBJ	84	205	185	125	241	56	170	188	209	157
OBJG	1	1	2	2	2	1	1	1	2	1

Table 4.1.: Unique verb lexemes in Kobalt subcorpora. Verb lexemes are counted by argument slot and do not sum up to the total number of unique verb lexemes per subcorpus: *Gehen* ('to go') in *Geht es der Jugend besser?* ('Are young people better off?') is counted once for SUBJ, OBJD, PRED each.

dep	CH-95	CH-115	CH-130	CH-160	L1	BEL-75	BEL-95	BEL-115	BEL-130	BEL-160
OBJA	168	307	245	163	233	98	331	304	289	211
OBJD	11	19	19	15	29	5	14	17	22	19
OBJP	54	109	101	98	152	22	109	98	96	77
PRED	29	62	52	35	57	15	72	68	66	40
SUBJ	116	226	189	141	242	59	194	198	198	139
OBJG	1	1	2	2	3	1	1	1	2	1

Table 4.2.: Unique argument lexemes in Kobalt subcorpora. Argument lexemes are counted by argument slot and do not sum up to the total number of unique argument lexemes per subcorpus: *Jugend* ('youth') in *Geht es der Jugend besser?* ('Are young people better off?') and in *Die Jugend ist (...)* ('young people are (...)') is counted once for OBJD and once for SUBJ.

and 4.2 show regular TTR and a transformed version of TTR normalized by the fourth root of the number of tokens, to make effects for equal text length more visible.

Most visibly in CH learners and BEL learners at around 600-700 tokens, learners of higher onDaF scores also show higher TTR for texts of equal length. Thus a higher lexical diversity as the result of a diversification with progressive acquisition is confirmed in this analysis, too. CH seems to score slightly higher in the transformed TTR compared to BEL, but lower than L1. Interestingly, there is still considerable overlap between groups, and also large variance in L1. In fact, the lowest-scoring L1 text lies lower in the transformed TTR than all of the higher scoring BEL-learners at the same text length.

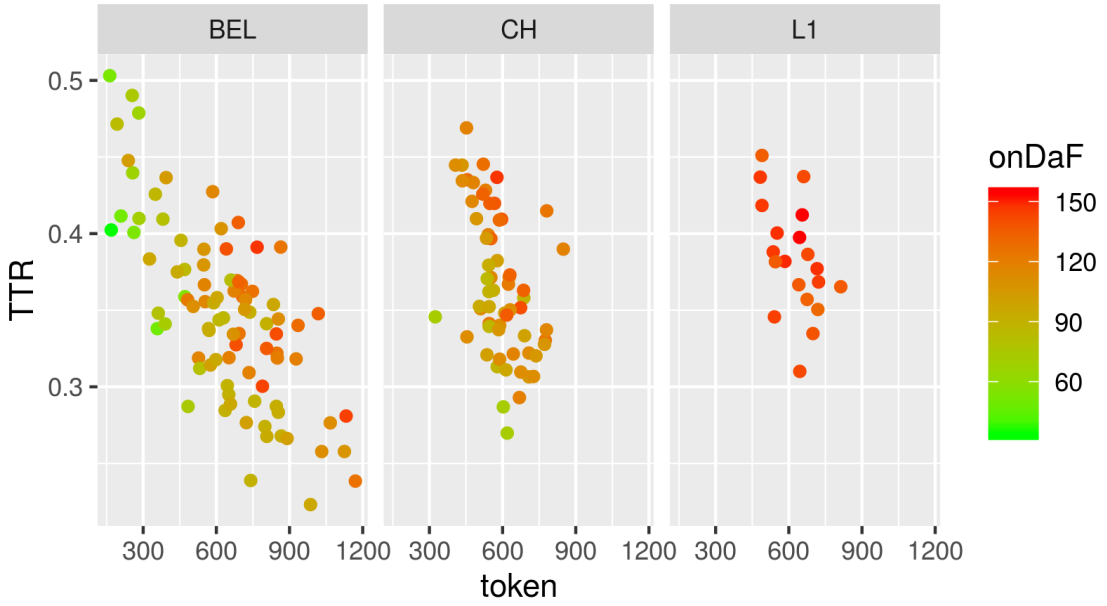


Figure 4.1.: Type-Token-Ratio in individual documents in Kobalt

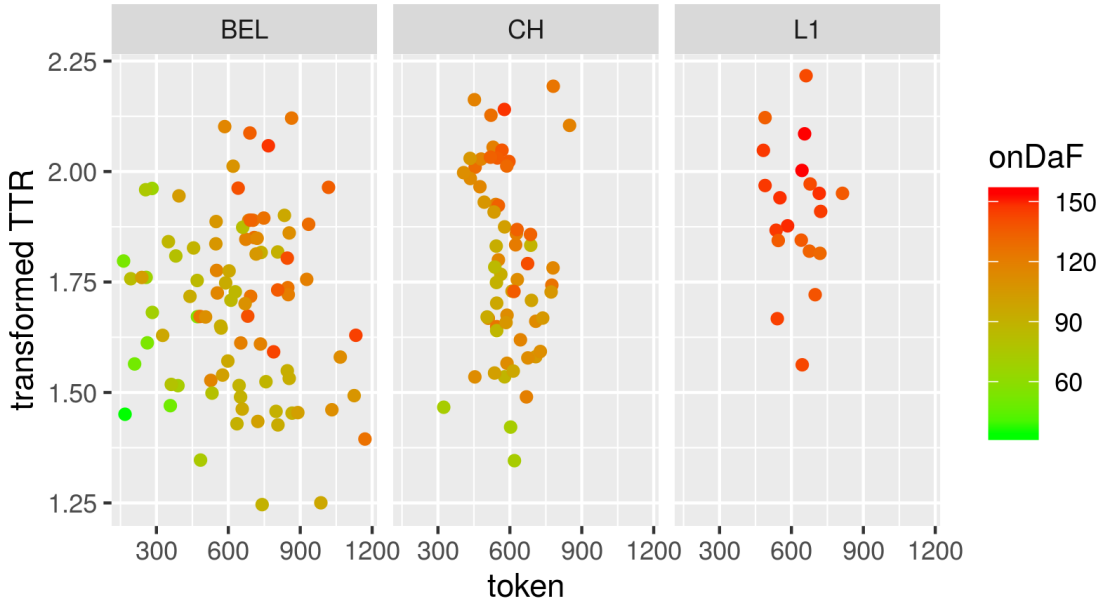


Figure 4.2.: Transformed Type-Token-Ratio ($TTR \cdot \sqrt[4]{token}$) in individual documents in Kobalt

This suggests that there is no unique 'target language-like' standard of lexical diversity that could be reached. It also suggests that inter-individual variance will be a relevant factor in the estimation of coselectional constraint.

4.1.2. Coselectional diversity

Tab. 4.3 shows the number of unique coselections. L1, positioned at the center, has a higher ratio of unique SUBJ to OBJA coselections, as was predicted. The same is not true of any BEL-corpora and CH-95 and CH-115, thus suggesting differences in the use of SUBJ and OBJA in L1 and L2.

dep	CH-95	CH-115	CH-130	CH-160	L1	BEL-75	BEL-95	BEL-115	BEL-130	BEL-160
OBJA	224	516	372	227	336	129	611	526	517	327
OBJD	17	32	30	21	49	7	22	23	27	30
OBJP	65	145	124	110	192	23	154	124	142	96
PRED	30	65	53	35	60	15	72	73	66	41
SUBJ	222	552	422	268	506	124	538	519	518	331
OBJG	1	1	2	2	3	1	1	1	2	1

Table 4.3.: Unique verb + dependency type coselections in Kobalt by subcorpora. Coselections are counted by slot, not by lexeme. If an argument is chosen as OBJA once and as OBJD the next time, it will be counted once for each of the two slots in this table.

For coselections, a diversification similar to the lexical diversification has not been hypothesized. Instead, it was expected that learners drop in coselectional constraint towards intermediate stages, then go back to higher levels of constraint at advanced stages. This is confirmed for OBJA, OBJP, and SUBJ in the ratio of unique combinations to all combinations used in that slot (in analogy to TTR, a type-coselection-ratio, fig. 4.1.2). Some observations:

- In the final onDaF-group, learners have higher unique coselection rates than L1 in all slots except SUBJ. This suggests that native speakers are more repetitive or less specialized in their writing compared to very advanced learners. It could also be an expression of a lack of cohesion.
- Up until this point, learners have lower unique coselection rates compared to L1, suggesting they are more repetitive. A corpus size or text length effect might be conflated here (like in the TTR in the lexeme analysis), because the lowest OBJA data point in BEL-95 correlates with the largest subcorpus. Since curves are distinctly different for the four slots though, this is unlikely to be the only explanation of the curve.
- Interestingly, OBJA and OBJP do not differ largely in terms of their relative uniqueness ratio in CH, but they do in BEL.
- PRED and SUBJ are more repetitive than OBJA and OBJP. This is incosistent with the hypotheses that SUBJ should be the least restrictive – here, OBJA and

OBJP take the most unique arguments, while SUBJ and PRED repeat arguments, i.e. take fewer unique arguments per slot. This is, however, to an extent a statistical artifact of the closed classes of copula verbs on the one hand (*sein, werden* 'to be, to become') and the frequent occurrence of pronouns in the SUBJ slot on the other hand. Thus it is a reflection of combinatorics, rather than a clearly linguistically determined effect.[2]

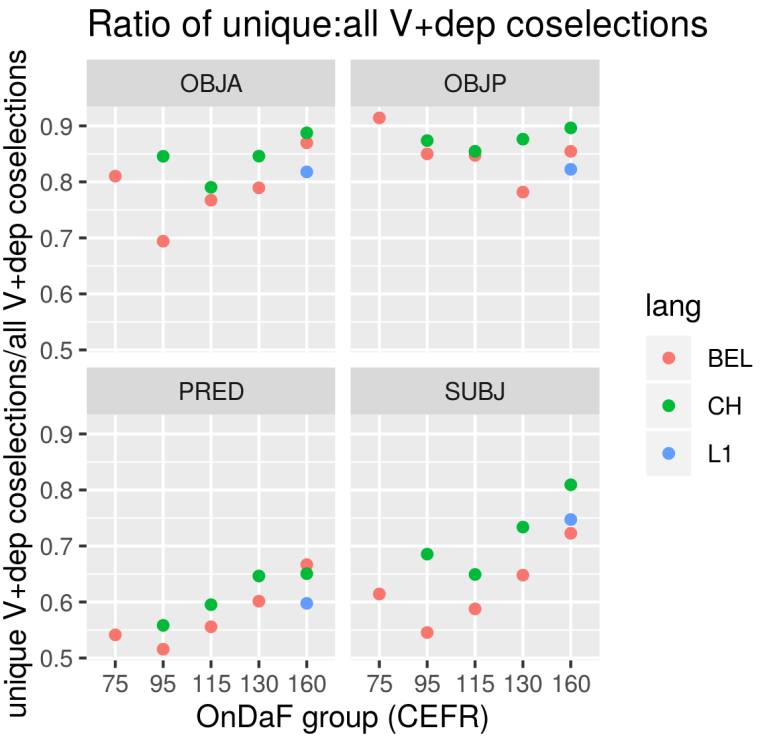


Figure 4.3.: Ratio of unique V + dep combinations divided by all combinations of that slot (in analogy to TTR, but for coselections). A u-shaped development is visible for OBJA with low points in BEL-95 and CH-115; for OBJP with low points in CH-115 and BEL-130, for PRED in BEL-95, and for SUBJ in BEL-95.

4.1.3. Morphosyntactic diversity of verbs

Fig. 4.4 gives an impression of the developments visible from verb category annotations alone, where verb categories are normalized against the total number of verbs in a document: Learners consistently overuse copula verbs compared to L1, which also suggests an overuse of predicates, and simplex lexical verbs (as opposed to particle or prefix verbs) at the lower acquisition stages. Their use of particle and particularly prefix verbs grows over

[2]Pronouns are included in the analysis here. From a statistical perspective, this may not have been an ideal choice, because pronouns are frequent and thus skew the statistics against less frequent lexemes. At the same time, linguistically speaking, verbs may possess coselectional preferences towards some pronouns over others, or towards pronouns over nouns, etc. Thus, if focused on form only, the model can include pronouns. This is in fact suggested by Römer et al. (2014), who reports pronoun preferences in argument slots of seemingly neutral verbs, like *to hear*. In a more semantically-guided model, pronouns should be dealt with separately. This remains for future research.

time.[3] Auxiliaries are also underused in most L2 corpora compared to L1, suggesting a different treatment of TAM.

As in the previous analysis, large variance exists in L1 texts, too: For simplex lexical verbs, their use ranges between less than 20% and over half of the verbs in a text; for modal verbs, the L1 texts cover a range from almost zero to 20%; category cx (constructions) also covers between zero and 10% in L1.[4] This means that 'target-like' baselines for comparison are not easily defined: A learner using zero constructions is as native-like in this respect as a learner whose every tenth verb is part of a construction.

In summary, verb category annotations in Kobalt point towards three tendencies:

- a diversification of vocabulary, and increasing lexical complexity of verbs, which, according to Plank (1984), should lead to higher coselectional constraints,
- a diversification of syntactic constructions (higher variance in cx), and
- growing syntactic complexity, particularly in the TAM domain (slight growth of auxiliaries and modals) and through exchange of copula-predicate structures for perhaps more complex syntactic structures.

Concerning the issue of coselectional preferences, the increase in complex verbs (particle, prefix) and decrease of simplex lexical verbs (last row in 4.4 of particular interest. While it is further evidence for an increasing diversification with growing onDaF, it also points towards a problem in the further analysis:

If learners use twice as many prefix verbs at later stages compared to earlier ones, and fewer particle verbs than L1 throughout the corpora, then that also means they use *different lexemes*. But if they use different lexemes from one another and from the L1 group, then tracking *the same lexemes* over time becomes a difficult endeavor, because

> (1) only the BEL group also writes increasingly longer texts, meaning that for them the different ratios might still translate to equal numbers (texts might just be filled up with prefix and particle verbs), but for CH, absolute frequencies are bound to decrease relative to their category's ratio, quantitatively suggesting higher or lower forces of attraction that may or may not be in accordance with the linguistic model; and
>
> (2) many verbs that exist in L1, which was supposed to be used as a baseline or target situation, will not appear in learner texts until rather late acquisition stages, meaning they cannot be tracked earlier.

[3] The final underuse of particle, but not prefix verbs corroborates results from a study of complex verb productivity in learners of German (Lüdeling et al., 2017) in the Falko corpus (Reznicek et al., 2010).

[4] The category includes modal infinitives and reflexive constructions such as *Man diskutiert sich heiß* ('discussions are running hot', literally: 'one discusses oneself hot')' (CH_033), some frequent constructions like *halten (für), haben (zu)* ('to consider', literally: to hold (someone) for (something)'; 'to have to') and *gehen (um)* ('to be about', literally: 'to go around'). *Gehen_cx* for *geht es ihnen gut/schlecht/...* ('are they doing well/badly/...') has been categorized separately because it occurs in the prompt and is therefore much more frequently picked up by the participants in their writing than other constructions, while it at the same time seemed relevant to be able to tell it apart from other uses of *gehen* ('to go'). See section 3.2.2 for details.

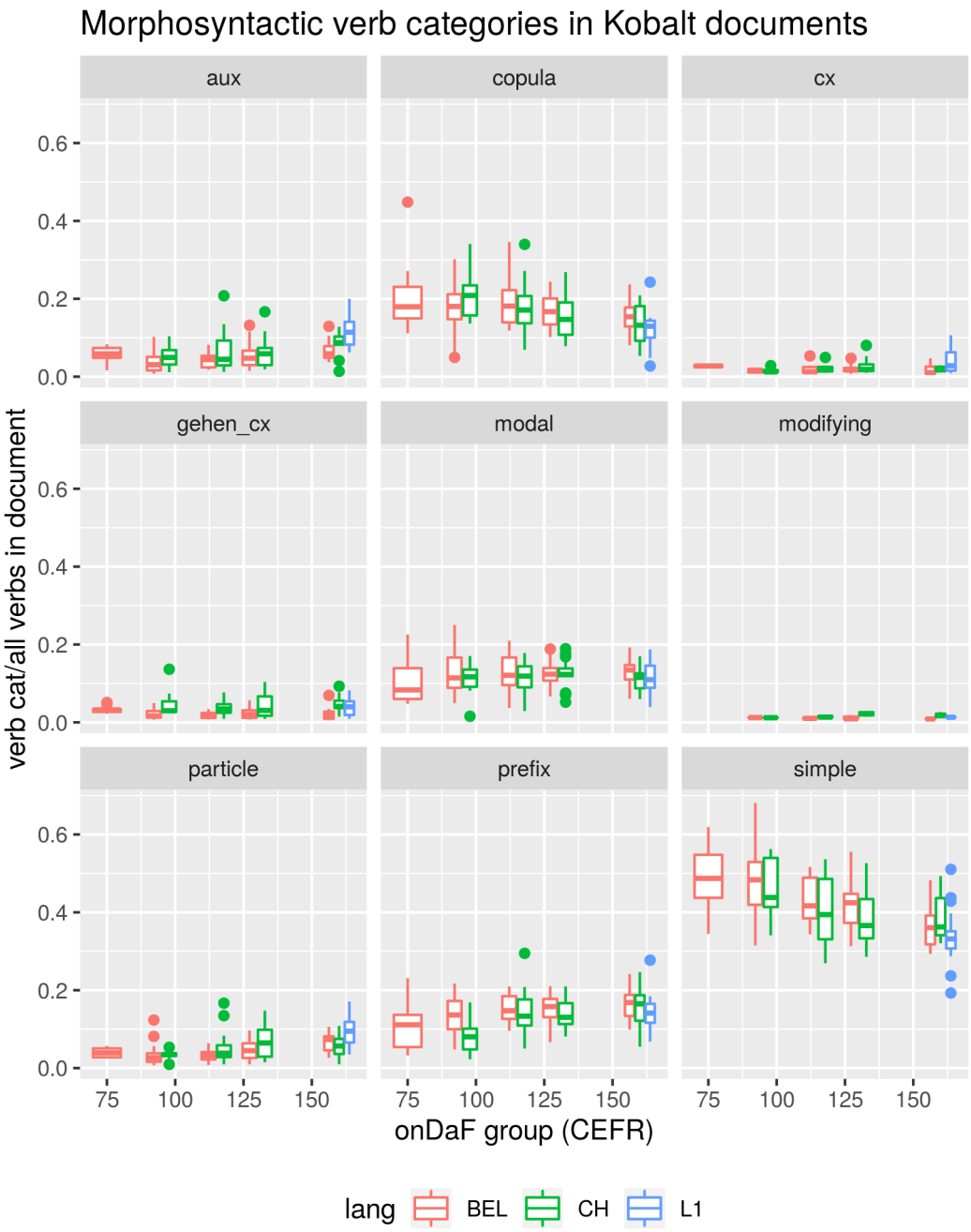


Figure 4.4.: Verb categories in Kobalt documents by onDaF group and language

4.1.4. Diversity in argument types

Figs. 4.5 and 4.6 give an overview of the VAS dependencies. Fig. 4.5 shows that for all groups, accusative (OBJA) and prepositional objects (OBJP), predicates (PRED), and subjects (SUBJ) are most frequent. Variance is generally higher at intermediate stages (more outliers), and higher for BEL than CH except for OBJP in CH-115. Frequencies are given per document and normalized against the number of finite verbs in each text.

Fig. 4.6 shows the same distributions on a free y-scale for better visibility of all devel-

opments. The largest and clearest development in relative frequency over time happens for prepositional objects, which are underused by learners of earlier acquisition stages compared to the most advanced learners and L1 group. This is consistent with previous observations from coselectional and lexical diversity in this chapter. Subjects are overused by both learner group and predicates are also slightly overused.[5]

Strikingly, variance in the use of all VAS is so high that all groups overlap to a large degree. While there is obvious growth in the use of OBJP, even the lowest-scoring learners with an average number of realized OBJP are within native speaker ranges. Genitive objects (OBJG) are the rarest among arguments, there are only 27 in the whole Kobalt corpus, and 21 of those are *der Meinung sein* 'to be of the opinion'. Thus their development should not be overrated from the plot.

For most dependency types (OBJA, OBJC, OBJD, OBJP, SUBJ, OBJG), CH and L1 lie closer to one another than BEL and L1. This is less expressed in OBJC and is not the case in the other clausal complements (SUBJC, OBJI). For those, however, the distribution in L1 spans most of the learner distribution, which might mean that those categories are more sensitive to stylistic choices.

[5] A lower ratio of realized subjects may stem from either coordination or from more clausal embedding in which the subject is not always realized (*Sie ging nach Hause, um ihren Hund zu füttern*, 'she went home to feed her dog', where the semantic subject of 'feed', 'she', cannot be realized in the infinitive clause), see also next subsection. Subjects for coordinated verbs were not reconstructed in this analysis. This would be desirable for future extensions.

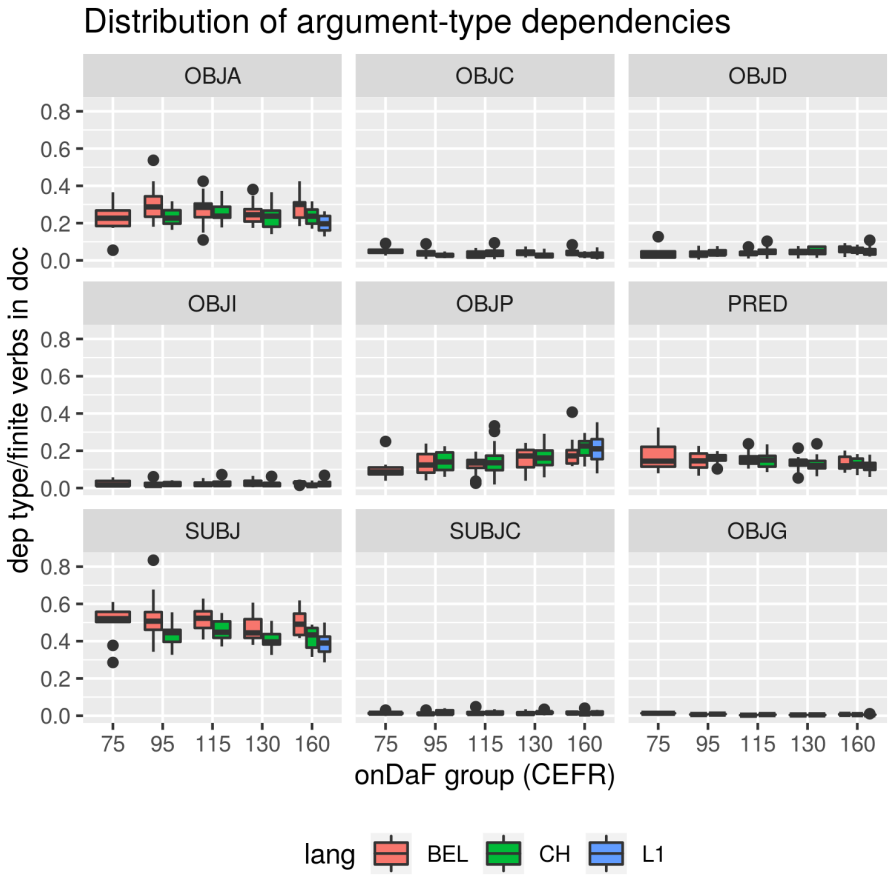


Figure 4.5.: Relative frequencies of verb dependents in individual documents

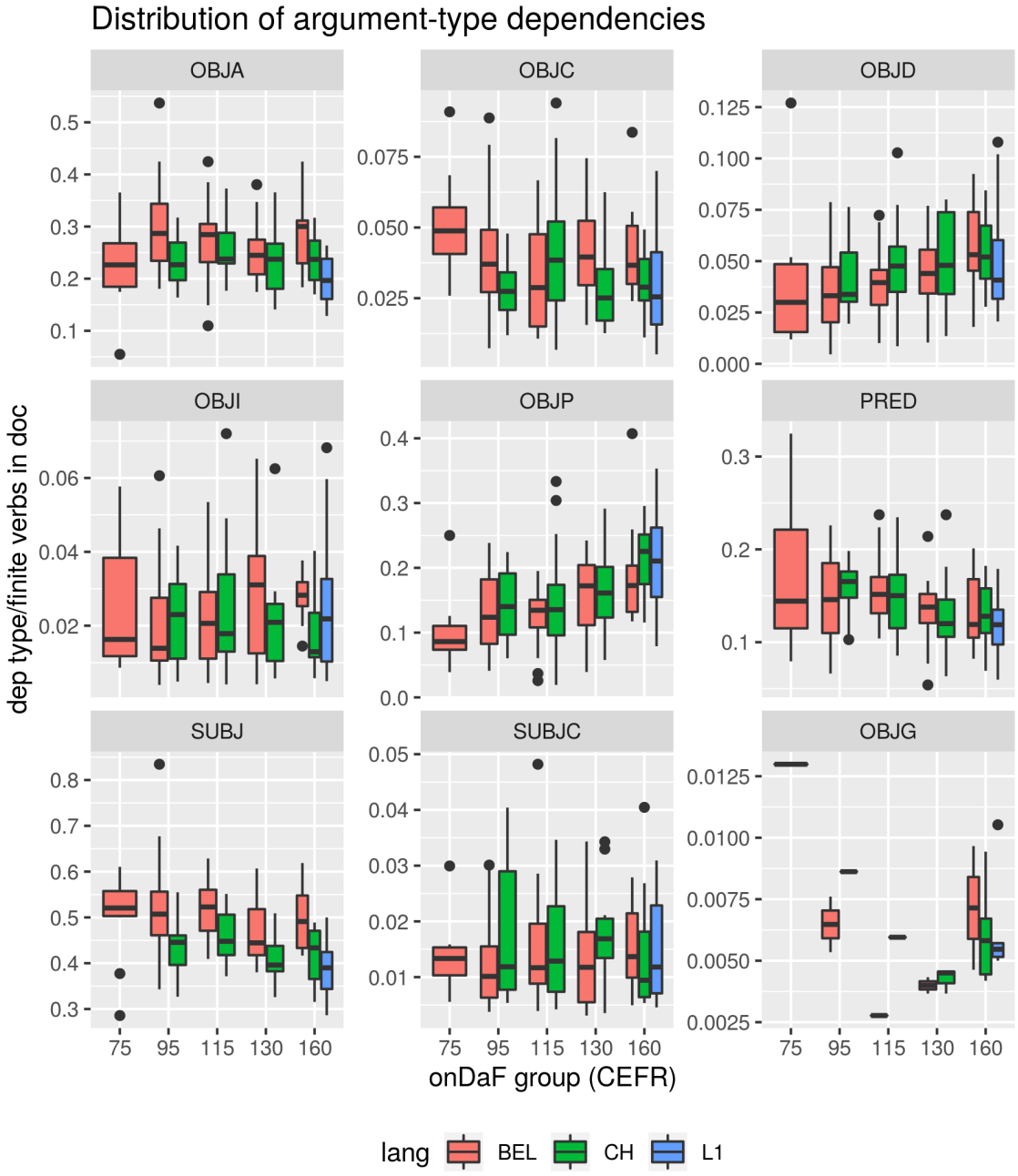


Figure 4.6.: Relative frequencies of verb dependents in individual documents, free y-scale

4.1.5. Distribution of verb-argument structures (VAS)

To gain an overview of the diversity of verb-argument structures (VAS) or subcategorizations as they are used in the Kobalt subcorpora, each verb is considered with all of its dependents. Each dependency type is only counted once,[6] and subjects are disregarded. This is because subjects can only be assigned to one verb in dependency grammar, and learners in particular like to coordinate verbs to long lists,[7] and because passives, deverbal adjectives that realize a VAS, clausal complements, and questions do not always realize subjects. It was considered that two VAS should not be counted as different only due differences in the clausal embedding of the head verb. Labels are then listed alphabetically, not in actual order of occurrence, and normalized to unique strings.[8] Thus, *Sie schreiben gerne Nachrichten, aber keine Briefe* ('They like to write messages, but no letters') is counted as two OBJA structures, 'write messages', 'write letters' (please note the overlap in labels which were previously used for dependency types rather than VAS – this is only the case for one-argument-VAS), and *Ihnen stehen viel mehr technische Geräte zur Verfügung* ('They have many more technical gadgets at their disposal') is counted as an OBJD_OBJP structure.

Fig. 4.7 shows the percentage of each verb-argument structure in the Kobalt subcorpora. For better legibility, fig. 4.7 only contains arguments structures that make up 2% or more of the total number of VAS in each subcorpus. A full overview can be found in the zenodo repository (`10.5281/zenodo.3584091`). As can be seen from the plot, simple structures dominate the picture across subcorpora and language groups:

- In all subcorpora, OBJA is the most frequent VAS that makes up at least half (L1) or up to 70% of all structures (BEL-095).
- PRED, which here marks either predicates linked to a subject with the copula *sein* ('to be') or non-NP arguments linked to construction slots (*es geht ihnen gut* ('they are doing well') is an instance of OBJD_PRED). Notably, predicates and accusative objects are overused by learners and are used less with increasing onDaF.
- Prepositional object-only structures, OBJP, are the next most frequently used VAS and mark the strongest growth from early to late acquisition stages as they did

[6] There are a few German verbs with two potential OBJA slots, like *nennen* ('to name') or *lehren* ('to teach'). These are labeled as OBJA2 in Foth's schema (Foth, 2006), but are not counted separately here, because they are very rare in Kobalt.

[7] Some examples:

- *Auf einer Seite bietet die moderne Technik von heute die Möglichkeit an, dass die jungen Leute mehr neue Meinungen erfahren, eine andere Kultur kennenlernen und ein vielseitiges Leben erleben können*, 'on the one hand, modern technology offers the opportunity that young people can learn about new opinions, meet another culture and live a more versatile life', (CH_051);
- *Sie schließen bloß die Tür, hören Rockmusik, surfen im Internet und machen unglaubliche und komische Dinge*, 'they just close the door, listen to rock music, surf the internet, and do incredible and strange things', (CH_058);
- *Jetzt hören wir sehr oft, wenn du etwas nicht kannst, hattest oder weißt, musst du diese Situation nicht zeigen*, 'now we hear very often, if you can't do, don't have or don't know something, you don't have to show this situation' (BY_084).

[8] This is therefore not a positional model that considers the order of arguments, which is not to say that this cannot also be relevant with respect to preferred expression. It is just not a practical distinction to make when numbers for the more complex argument structures are low anyway and the combinatorics would make them dwindle into next to nothing.

in the dependency type analysis. This is particularly relevant for the analysis, because according to Plank (1984), direct objects (OBJA) are the least coselectionally constrained out of the objects, while it must be assumed that OBJP are most lexicalized because this class contains a number of support verb constructions and is generally considered to be more non-compositional due to semantically bleached and lexicalized prepositional linking.

From this overview, it seems clear that a quantiative analysis of covarying lexemes like it was introduced in Stefanowitsch and Gries (2005) is not feasible given that there is only one somewhat frequent VAS that occurs in most subcorpora and has more than one argument, namely OBJA_OBJP, out of which OBJP is fixed and therefore nearly trivial in terms of coselection since word senses are bound to the chosen preposition. Ditransitive constructions are not particularly frequent in Kobalt and therefore cannot be used for quantitative analysis either.

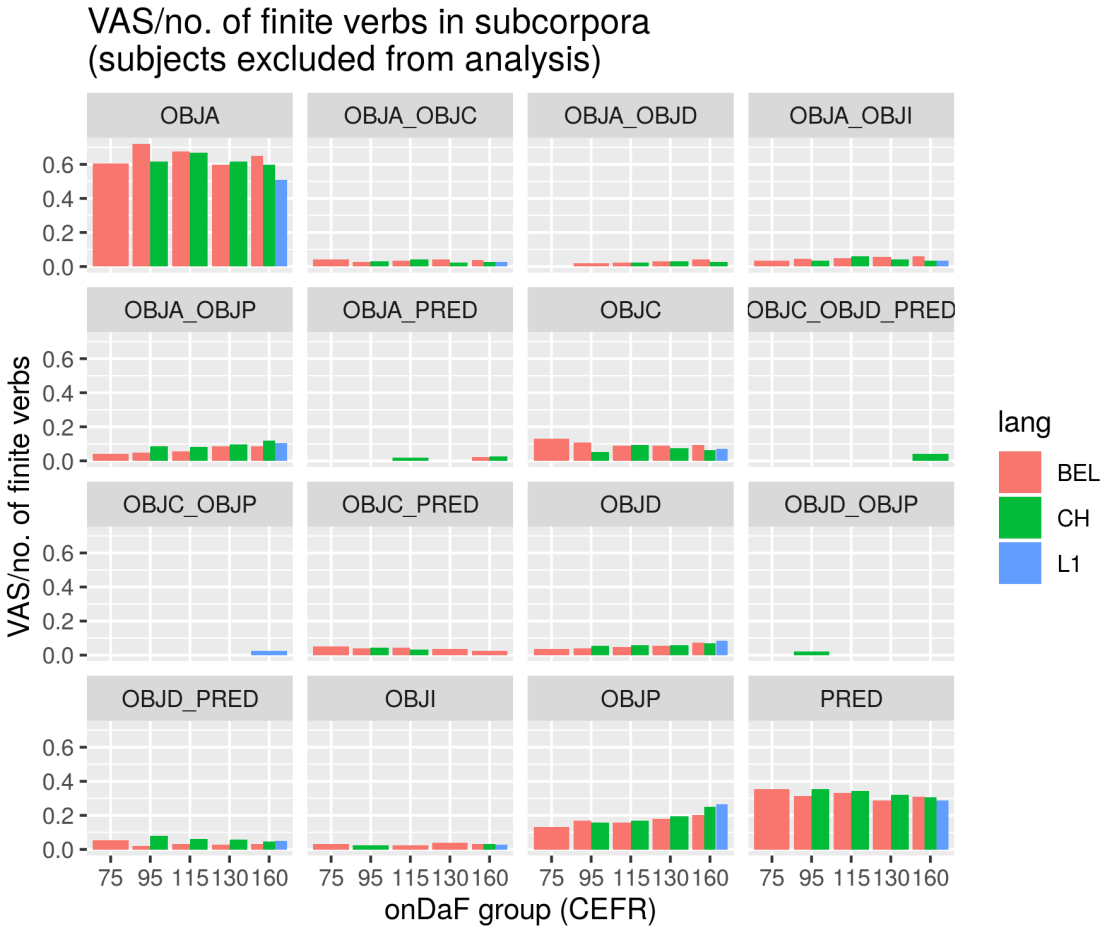


Figure 4.7.: Percentage of VAS in Kobalt relative to the number of finite verbs, subjects exluded. For better legibility, the plot only includes VAS that make up at least 2% of the total VAS. OBJD_PRED and OBJD occur at lower acquisition stages in BEL, too, but at lower percentages. A plot of the full distribution is included in the zenodo repository (`10.5281/zenodo.3584091`).

4.1.6. Summary: Diversity and Diversification

In summary, a process of diversification through the course of acquisition is confirmed on a lexical, and on a morphosyntactic level for verbs, and on a syntactic level through an increase of constructions and complex argument slots as well as more complex verb argument structures. Some of the changes also imply a specialization, such as the increase in complex verbs which express more fine-grained or specialized meanings than their simplex roots and the incrase in OBJP, as has been argued before.

4.2. Similarity

The next question is whether texts are overall similar enough to be compared. Similarity can exist on a number of linguistic levels, such as lexical, lexicosyntactic, semantic choices, but also register and topic. It has already be shown that there exists large variance between texts in all groups, but that general developments can still be seen. This section will show that there is considerable similarity in core aspects, namely part-of-speech (POS) distribution, vocabulary, and most frequent verb and argument lexemes. It will then show that, despite this, similarity in coselections is limited.

4.2.1. Part-of-speech distributions

Fig. 4.8 reports the part-of-speech distribution in Kobalt subcorpora, where subcategories are grouped for better legibility.[9] Some differences in the distribution of POS tags between subcorpora do exist: Learners slightly underuse adverbs and adjectives in sum (the upper two color blocks reach slightly less low than in L1), which is consistent with Hirschmann et al. (2013) and Hirschmann (2015) who also report underuse of modifiers in learners vs. native speakers; And that it seems that the distribution in L2 is particularly receptive to changes in the domain of pronouns and determiners. At the same time, the distributions both between L2 and L1 and between subcorpora in L2 are also remarkably similar. Later analyses will show that texts are still also remarkably different, which in combination suggests that POS is significantly more structural and underlies much less choice than some of the other aspects discussed above and below. This will be referred to later in this chapter.

4.2.2. Shared vocabulary

When speaking of coselectional *constraints* and idiomaticity, the underlying assumption is that there is lexical overlap, otherwise the same items could not be coselected. This section shows the degree to which lexical sets between texts overlap in order to gain a better understanding of how much similarity in coselectional constraint can be expected. Fig. 4.9 shows a heatmap of lexeme overlap in L1. The numbers signify the percentage of the lexicon of the text on the x-axis that is covered by the the overlap with the text on the y-axis. All lexemes including function words are included in this analysis. For example, DEU_003 (x-axis) and DEU_021 (y-axis) share 90 lexemes. For DEU_003, this makes up 40.4% of its lexical inventory, but DEU_021 (on the x-axis) is more diverse, so that the same 90 shared lexemes make up only 30.3% of its lexicon. Since each text shares all

[9] The full distribution can be found in repository (`10.5281/zenodo.3584091`). It should be used with caution, since TreeTagger (Schmid, 1995) does not reliably distinguish between fine-grained subcategories, and tags have only been corrected for verbs.

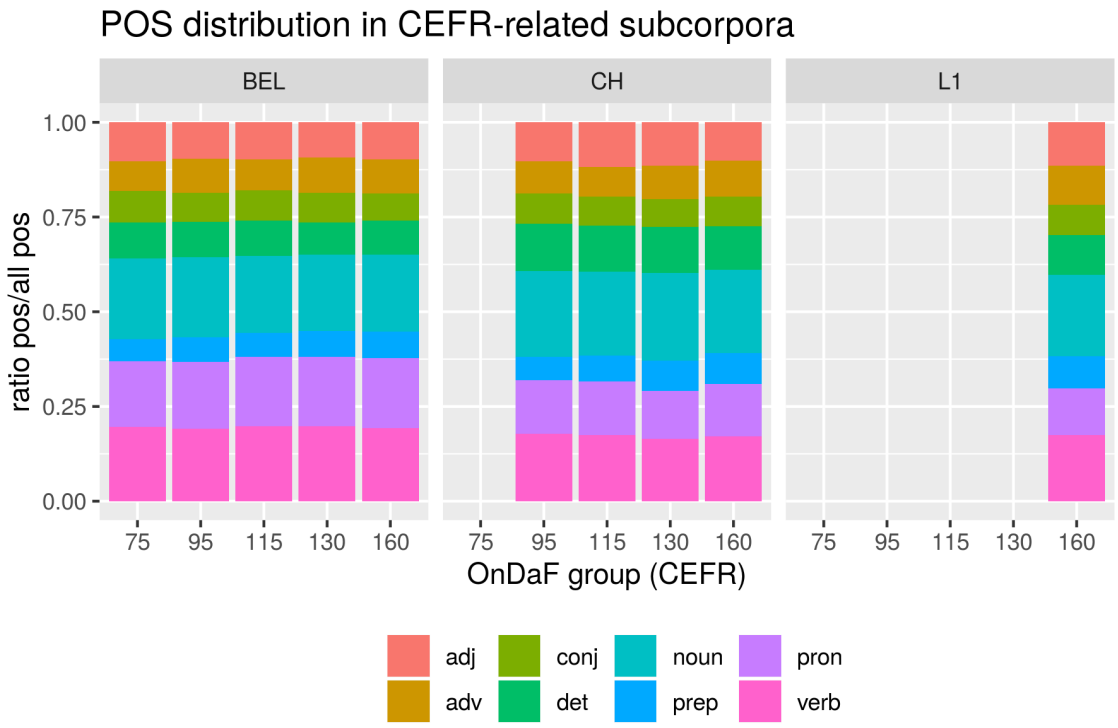


Figure 4.8.: Distribution of POS categories in Kobalt by subcorpora and language

of its lexemes with itself, the diagonal is always 100. Several interesting observations can be made here:

- The percentage that shared lexemes cover between L1 texts varies between 25 and just under 50%. That is quite the variance considering that L1 texts in Kobalt are relatively similar in length and topic.
- High overlap may stem from either high lexical similarity (texts 1 and 2 contain the same lexemes) or from a large difference in lexical diversity (text 1 contains the same lexemes as text 2 and many more). It appears from this plot that both phenomena exist, because there are on the one hand rows and columns that are darker, indicating that a single text is either lexically covered to a large extent by many other texts (darker columns) or that a single text convers many other texts (darker rows), and since the sets of all texts are not similar (otherwise there would not be such a high variance), it must mean that it covers different texts in different ways.
- However, there are also isolated cells that are darker, speaking for individual similarity as in DEU_005 (x-axis) and DEU_011 (y-axis), where 49.3% of DEU_005 and 46.4% of DEU_011 are covered by the same overlap. There is also one text that appears different from the others, because it is covered to a much lesser degree by the overlap with other texts, and that is DEU_020 (on the x-axis). At the same time, it does not seem to cover large parts of the vocabulary of other texts (on the y-axis) either, unlike DEU_021, DEU_014, and DEU_015 do. This suggests that the text is not necessarily richer or more comprehensive in vocabulary, but simply different from the other ones. This is interesting, again, with respect to the notion

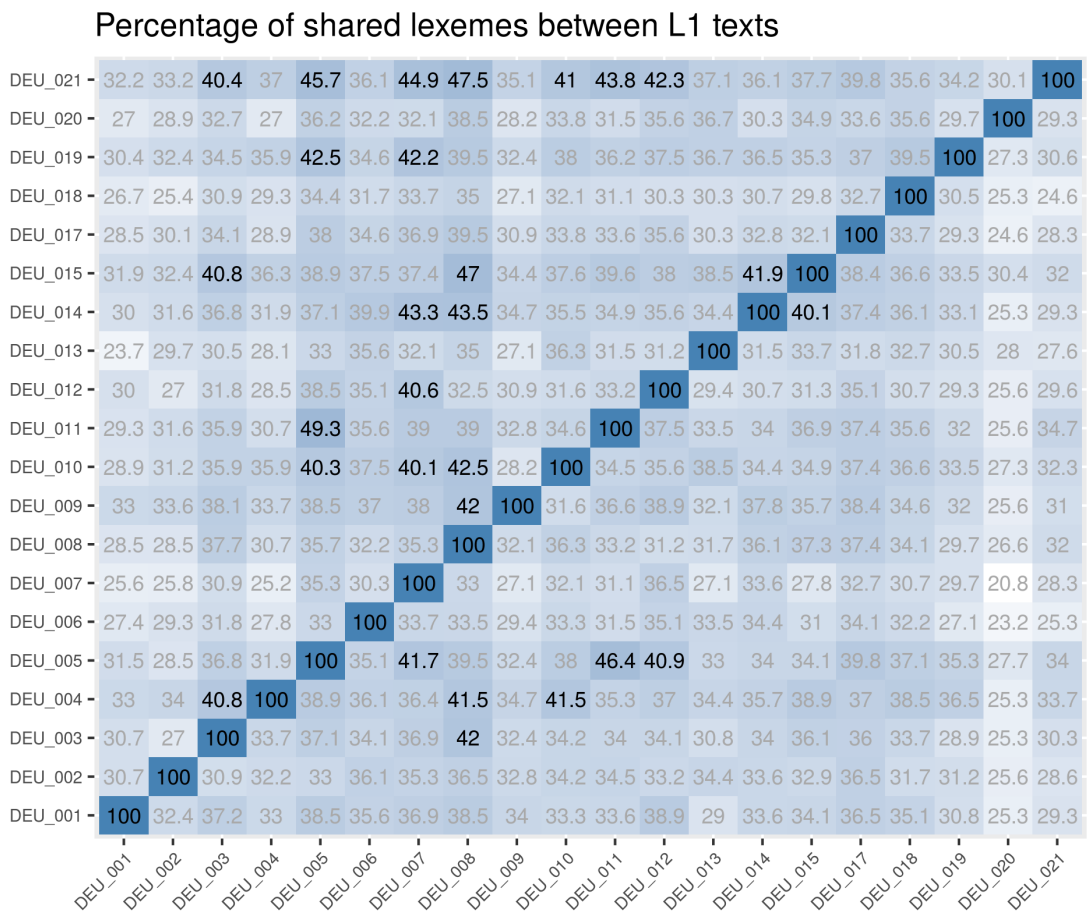


Figure 4.9.: Heatmap of lexeme overlap between texts in L1. Percentages are reported with respect to the text on the x-axis, i.e. the lexical overlap between a text on the x-axis and a text on the y-axis (for example 90 lexemes) covers the represented percentage of all lexemes in the text on the x-axis. Black font marks lexical coverage of $\geq 40\%$.

of a target language, where native speakers are presumed to be a somewhat homogeneous group that learners need to adjust their linguistic behavior to. Here, some native speakers are rather homogeneous with one another, while others are not, and yet others subsume those two groups in overall higher lexical diversity (share many lexemes and add many more).

How similar are learners in that respect? Since there are many more learner texts and the matrix grows quadratically, their heatmaps are less printing format-friendly, especially with the numbers staying legible. Therefore, only simplified heatmaps will be presented here, where colors indicate values above a certain threshold and discuss grouped results further below. Full heatmaps can be found in zenodo (`10.5281/zenodo.3584091`).

Fig. 4.10 shows that relative overlap is much higher in BEL learners. In the left plot, matrix cells are marked in blue for values of 40% or higher, which was a value only few text combinations in L1 reach. For BEL, almost half of the matrix is marked blue, and in particular the first quarter or so to the left. These are texts of low onDaF ranges, too,

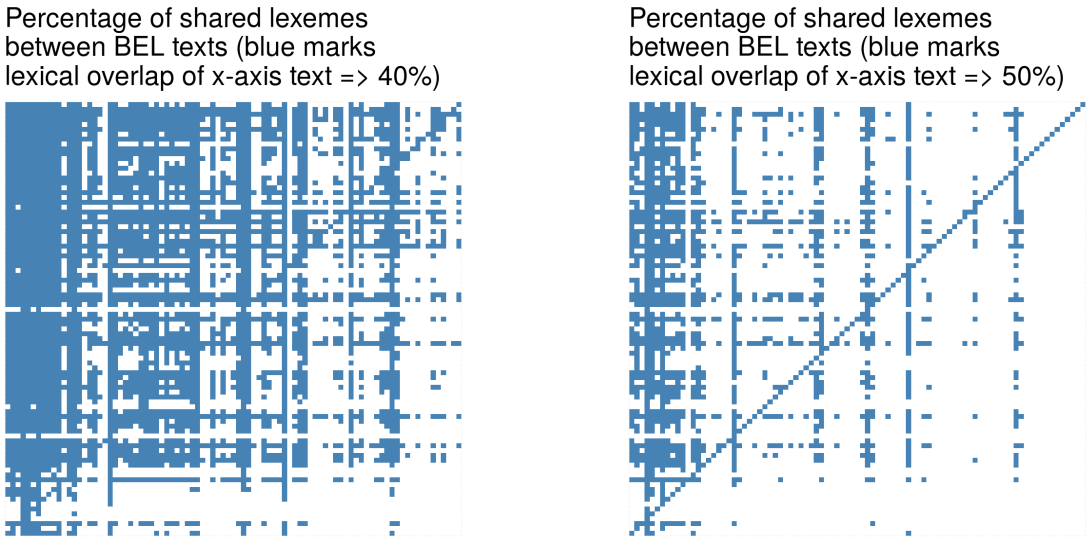


Figure 4.10.: Simplified heatmaps for lexical overlap between BEL texts, blue fields indicate an overlap of $\geq 40\%$ (left) and $\geq 50\%$ (right)

which makes sense, because more advanced texts will likely cover more of the lexicon of an early intermediate text and then go beyond, but not vice versa. This is also a text length effect, but 50% are still reached by a large number of texts across onDaF ranges and thus text lengths. It appears that despite the large variance in onDaF scores and text length, BEL learners' writing is more similar within-group than L1 writing in terms of lexical choice. This might be related to stylistic or register choices, or it might be a characteristic of learner language, or of BEL-L1 German-target language interlanguage specifically.[10]

Looking into the CH learners (fig. 4.11), it seems that there is indeed an SLA-specific effect, where 40% are reached by many more texts pairs in CH, too, and 45% still by more than the 40% in L1. CH therefore in this analysis seems to fit in between BEL and L1, still suggesting that a general SLA effect is at play, but works differently for the two languages. Plots of the other language combinations can be found in the zenodo repository (BEL vs. CH, CH vs. BEL, `10.5281/zenodo.3584091`).

This is also confirmed in grouped results. Fig. 4.12 shows that

- indeed, for each language group, the highest average lexical coverage of a text is reached in pairs from the same language group, and the same is true for maximum lexical coverage;
- L1 is less covered by both learner groups, while overlap with L1 does cover L2 texts about as well as the other L2 group respectively.

This can be interpreted as meaning that learners use vocabulary similarly to L1 to a degree, but L1 goes beyond what learners do, which is in line with the predictions. At the

[10] It might also be a teaching effect. However, in that case, one would expect clusters by onDaF to be visible: Students are taught in classes and are most likely to retain vocabulary that was learned shortly before, and this will be most similar in the same class even when the same teaching material is used every year. Such an effect is not strikingly detectable. It may still exist, but be masked by the many lexemes that could potentially be affected.

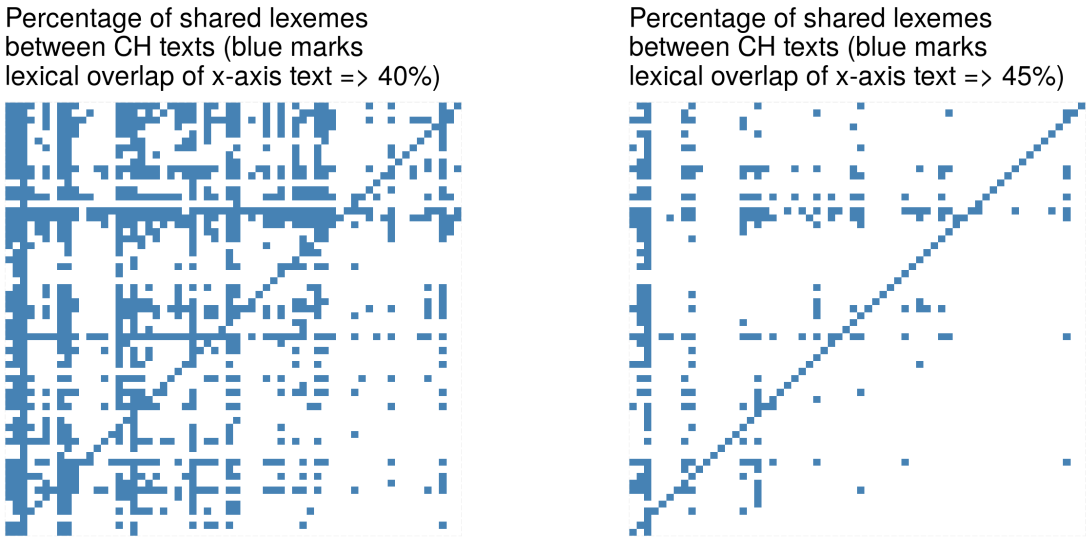


Figure 4.11.: Simplified heatmaps for lexical overlap between CH texts, blue fields indicate an overlap of $\geq 40\%$ (left) and $\geq 45\%$ (right)

same time, learner groups between one another cover less lexical material than within each group, but still more compared to L1. In other words, L1 usage predicts L2 usage better than vice versa, and either L2 predicts more for L2 than for L1, i.e. there is a language group effect, but a stronger L2 vs. L1 effect.

- In addition, BEL is covered better by CH and CH is covered equally well by either.

Interestingly though, for the minimum, this picture changes:

- L1 overlap covers all three language groups at a much higher minimum rate, namely at 22.5% vs. less than 15%;
- BEL texts have a lower minimum coverage for BEL texts vs. CH texts, meaning that in the minimum case, more of a CH text vocabulary vs. a BEL text vocabulary is covered or predicted from a BEL text.

It seems than that L1 does have an idiomaticity or perhaps a language structural effect that becomes visible through lexical overlap and that is defined through a lower bound of overlapping vocabulary at some 20% between texts of this kind and size; while learners may deviate to a lower overlap. Some of the effects for BEL here may stem from text length. Still, since L1 predicts higher minimum overlap vs. BEL – despite the fact that there are more texts of each length in BEL (meaning that any text can find a more similar text within group than outside of the group) – this cannot be the only explanation. Coverage distributions divided by onDaF groups are included in the repository (`10.5281/zenodo.3584091`), but are inconclusive. No clear assimilation towards higher onDaF groups can be found. This may be due to the competing effects of diversification and approximation to native-like selection.

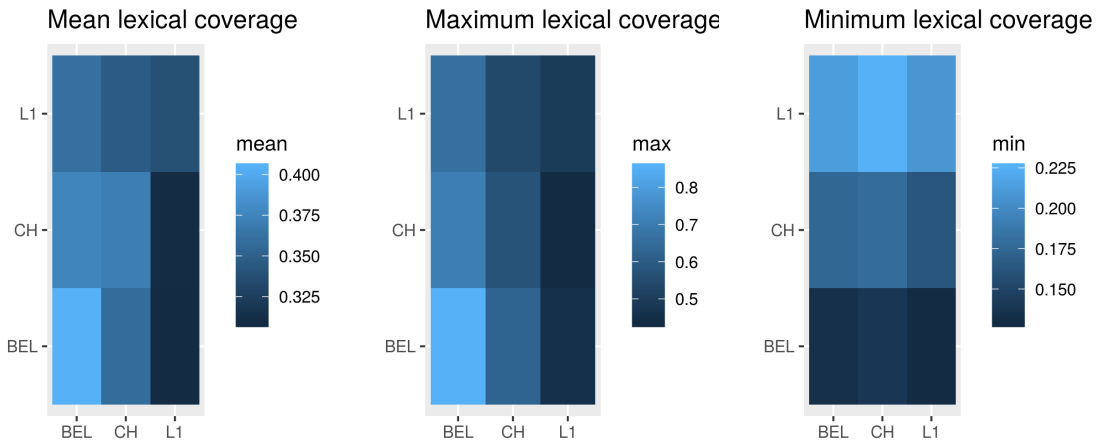


Figure 4.12.: Mean, maximum and minimum lexical coverage through intersection by language group

4.2.3. Most frequent verbs and arguments

Now that it has been shown that there is a considerable amount of shared vocabulary, particularly between learners, the question remains of whether lexemes are distributed in different ways regarding their frequency of occurrence. Figs. 4.13 – 4.15 show the twenty most frequent verbs in all Kobalt subcorpora. There are several findings worth discussing here:

- For all subcorpora, *sein* ('to be') and *haben* ('to have') are the most frequent verbs, which is expected since they are functionally diverse (as auxiliary, construction, copula, lexical verbs) and obligatory in many syntactic contexts of German.
- In L1, following those in rank is *werden* ('to become', future of *sein*, and auxiliary in future tense and passive), which is lower ranked in all L2 corpora, but rises in rank towards higher onDaF ranges.
- The next ones in L1 are *gehen, können, müssen* and *geben* ('to go', 'can', 'must', 'to give'). All of these can be seen as functional, two more clearly due to being modals; *gehen* because it is included in the prompt in a constructional sense *gut/schlecht gehen* 'to be well/unwell'; and *geben* because it mostly occurs not as the distransitive 'give' but as an existential *es gibt/es gab* ('there is/there was').
- On ranks 8, 9, and 10 in L1 are three more lexical, but still semantically light verbs: *kommen, leben* and *machen*, 'to come', 'to live', and 'to make'.

Interestingly, *kommen* is not included in the top 10 most frequent verb list in any of the L2 subcorpora, *leben* only appears in the list in four out of nine L2 subcorpora, while *machen* tends to be overused in L2, appearing three or more ranks higher in five out of nine L2 subcorpora.

Verbs that are not included in the top 10 most frequent verbs in L1, but do appear in the lists of the L2 most frequent are mostly idiosyncratic to one or two L2 subcorpora. An exception is *sagen* ('to say'), which is highly ranked in all but one of the L2 subcorpora, but only appears in the L1 list on rank 19, which is a half a magnitude smaller with a an

absolute count of 11 occurrences in L1 vs. over 50 in subcorpora of comparable size in BEL. In summary, *machen* and *sagen* ('to do/to make', 'to say') are overused by learners compared to L1 in Kobalt, and *leben, kommen, werden* and *müssen* ('to live, to come, to become', 'must') are underused by learners.

For ranks 11-20,

- most shared verbs between L1 and L2 are modals *mögen, wollen, sollen* ('may', 'want', 'shall, should') or relate to discursive or epistemic orientation (*sagen, denken* 'to say', 'to think');
- With respect to the latter, there are several more that appear in the list for L2, but not L1: *meinen* 'to mean', *wissen* ('to know'), *finden* ('to find'), *glauben* ('to believe').
- Learners appear to cluster in language groups to a degree, although there is some overlap (BEL: *sehen, meinen, wissen, helfen, finden, glauben, studieren* ('to see', 'to mean', 'to know', 'to help', 'to find', 'to believe', 'to study'); CH: *lernen, entwickeln, studieren, wissen* ('to learn', 'to develop','to study', 'to know'), pointing at differences in both topic and writing style.
- BEL-160 and CH-160 also each include verbs that are not highly ranked in any of the other subcorpora.

Fig. 4.14 shows the percentage each of the verbs makes up relative to all verbs in the respective subcorpus. An aspect that has been mentioned in the previous section is visible here too, namely that of lexical diversification: In the lower onDaF ranges, the higher ranked verbs make up a higher percentage of all verbs in the subcorpus, with over 20% of *sein* ('to be') in the lowest BEL and CH subcorpora vs. less than 17% in BEL-160 and less than 15% in L1. The lower-ranked verbs in the top 10 of the list still make up over 2% in the lower onDaF L2 corpora, about 1.5% in the upper onDaF L2 corpora, and just 1.08% in L1. This is consistent with the observation that learners are generally less productive in writing (Zeldes, 2013a; Lüdeling et al., 2017), rendering lexeme distributions 'more Zipfy'.

Fig. 4.15 gives the absolute frequencies for the most frequent verbs. As can be seen, frequencies drop to numbers of less than 25 quickly for verbs that are not mostly functional, and goes down to only five to eight absolute occurences for ranks 19–20 in the smallest subcorpora. This will become a relevant challenge in the next section and also means that beyond the first 10 to 15 ranks, the ranking in the smaller subcorpora especially is somewhat random because all occurrences of the last ranks are within the same order of magnitude. This will be discussed as a challenge in section 4.3.

Figs. 4.16–4.18 show a comparison of the percentage of verb argument lexemes undivided by slots as they appear in L1 ranks. Clausal or verbal arguments (OBJI, OBJC, SUBJC) are not considered in this analysis. For example, *Jugend* ('youth') is the most frequent argument in L1, it makes up 5.58% of all L1 arguments (OBJA, OBJD, OBJG, OBJP (noun complement to PP in OBJP), PRED, or SUBJ). It is also very frequent in the other subcorpora, and makes up 10.24% of the CH-95 subcorpus, but only 3.04% of the BEL-95 subcorpus.

Ranks > 8 or percentages $< 3.51\%$ mark a break in L1 both in terms of frequency (the next arguments are roughly half as frequent or less) and topic or source (except for *Kind*, 'child'). The most frequent arguments in L1 are either functional (reflexive pronoun

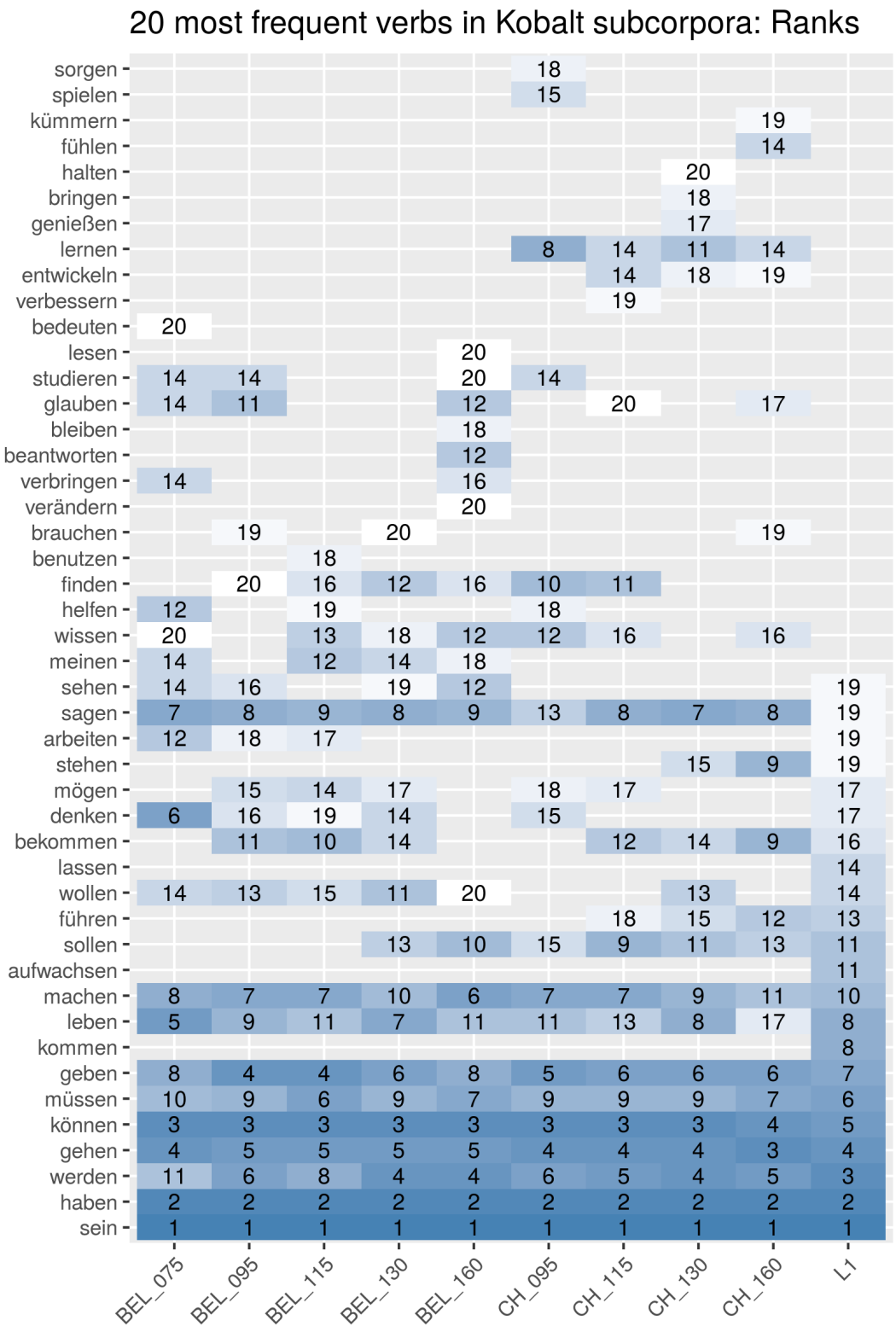


Figure 4.13.: 20 most frequent verbs in Kobalt. Darker colors indicate higher rank.

sich, definite article/demonstrative pronoun *d*, indefinite pronoun *man*, 'one') or copied from the prompt or closely related lexical material (*Jugend, gut, Generation, Jugendliche*, 'youth', 'good/well', 'generation', 'adolescents').

This figure shows that for the first 25 ranks in L1, lexical material between L1 and L2 is very similar both in terms of lexeme choice and distribution, although there are some interesting cases of over- and underuse, typically divided by learner groups, too. For example, learners consistently underuse the words *Vorteil* ('upside, advantage'), *Nachteil* ('downside, disadvantage'), and *Aussage* ('statement') compared to L1, but both groups overuse *Leben* ('life') and *Eltern* ('parents'), pointing perhaps towards different topics, but also registers of their writing. *Vor-/Nachteil* and *Aussage* are frequently used as discourse markers or for topic introduction in analytical registers of German writing. Functional *welch* ('which') and *dies* ('this/these') are underused by learners, but *was* ('what','which' in relative clauses) is overused, pointing towards differences in syntactic complexity and/or definiteness or specificity. On the whole, however, those first 25 ranks are mostly similar because they are made up from lexemes that are either closely prompt-related, functional, or rather unspecific (*Mensch*, 'human, man, person', *Möglichkeit* 'chance, possibility, opportunity, option').

Lexical overlap in arguments is still remarkably high both between L1 and L2 and the two L2-groups at ranks 26–50 (fig. 4.17) and 51–100 (fig. 4.18). There are some more visible preferences in L2 now, such as an overuse of *Zeit* ('time') in BEL and *Gesellschaft, Lebensstandard* ('society', 'living standard') in CH, but overall, if a lexeme is used in one of the learner corpora, it is most often also used in the other ones. For better orientation regarding absolute frequencies: Rank 1 in L1 (*Jugend*) occurs 100 times, while rank 50 in L1 occurs 6 times. Ranks are not unique by frequency here, which means that ranks 51–60 also belong to words that appear 6 times in L1.

Absolute frequencies after those ranks quickly dwindle down to hapaxes, where a comparison between an occurrence in one corpus and a non-occurrence in another seems more random. Overall, a comparison of argument lexemes shows a large overlap for the most frequent 50 lexemes. Despite some idiosyncrasies by language group for the ranks beyond that there is still an apparent agreement of the general lexicon to be used in response to the prompt. It will be shown later that, despite this, there are still remarkable differences in register and style, and that the number of identical coselections is not very high despite the large overlap.

Differences exist mainly where register-awareness may play a role: Learners use *finden, meinen, wissen* 'to find', 'to think (to have an opinion)', 'to know', while native speakers appear to avoid this kind of direct expression of personal thoughts and opinions. Instead they use more arguments that may be interpreted as belonging to a discourse-orienting meta level, like *Vorteil/Nachteil, Unterschied, Aspekt, Fall, Begriff, These* ('advantage/disadvantage', 'difference', 'aspect', 'case', 'notion', 'thesis, hypothesis').

Results can also be interpreted as confirming the prediction of the existence of a shared core lexicon relative to a topic, but that diversification in a corpus of this size leaves very little room in between shared lexemes across subcorpora and hapaxes. It is possible that the space in between, which may be filled with more fine-grained variations and language-group-specific preferences, require a larger dataset to fully unfold.

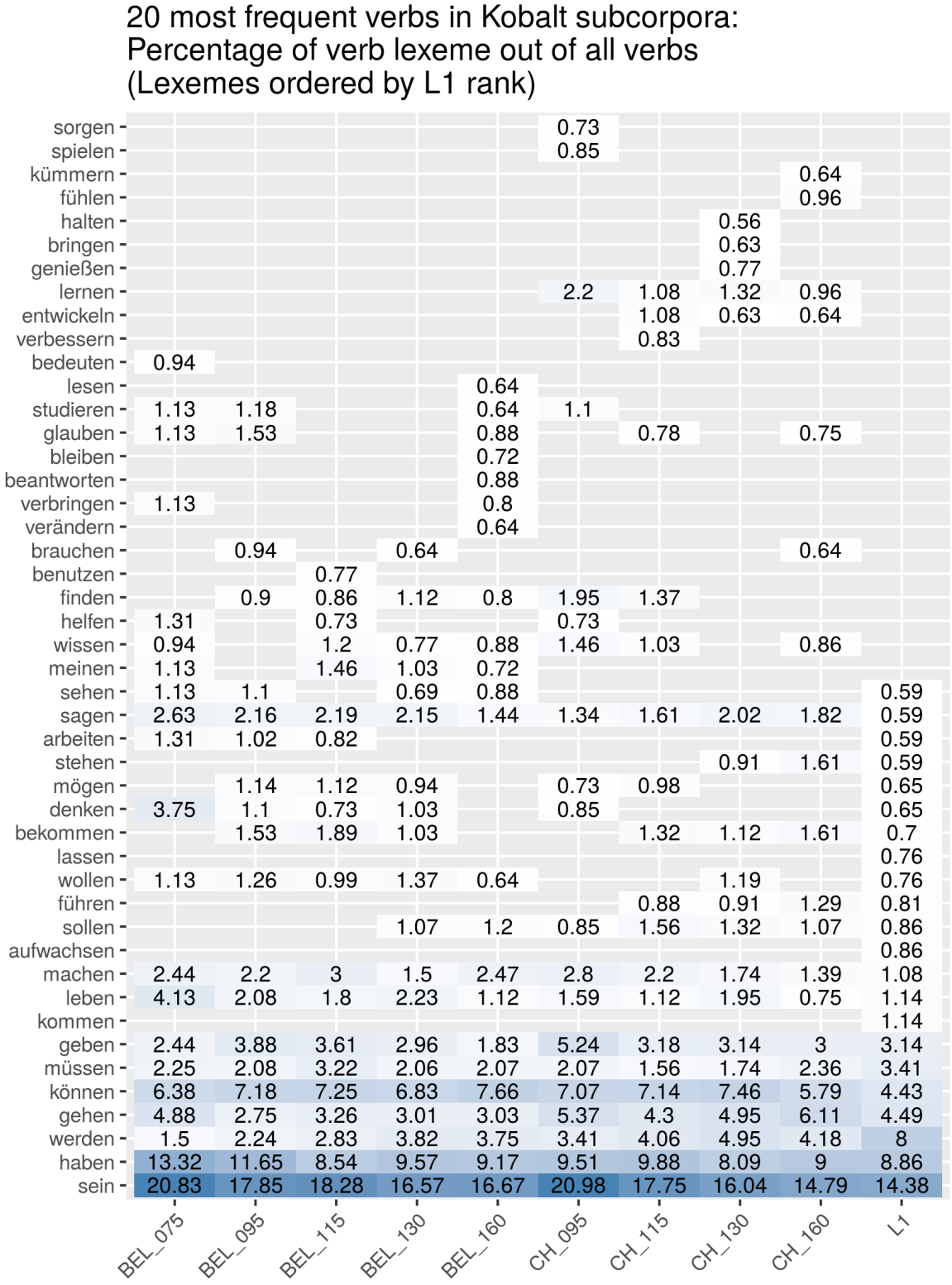


Figure 4.14.: Percentage of 20 most frequent out of all verbs in Kobalt subcorpora. Darker colors indicate higher percentage.

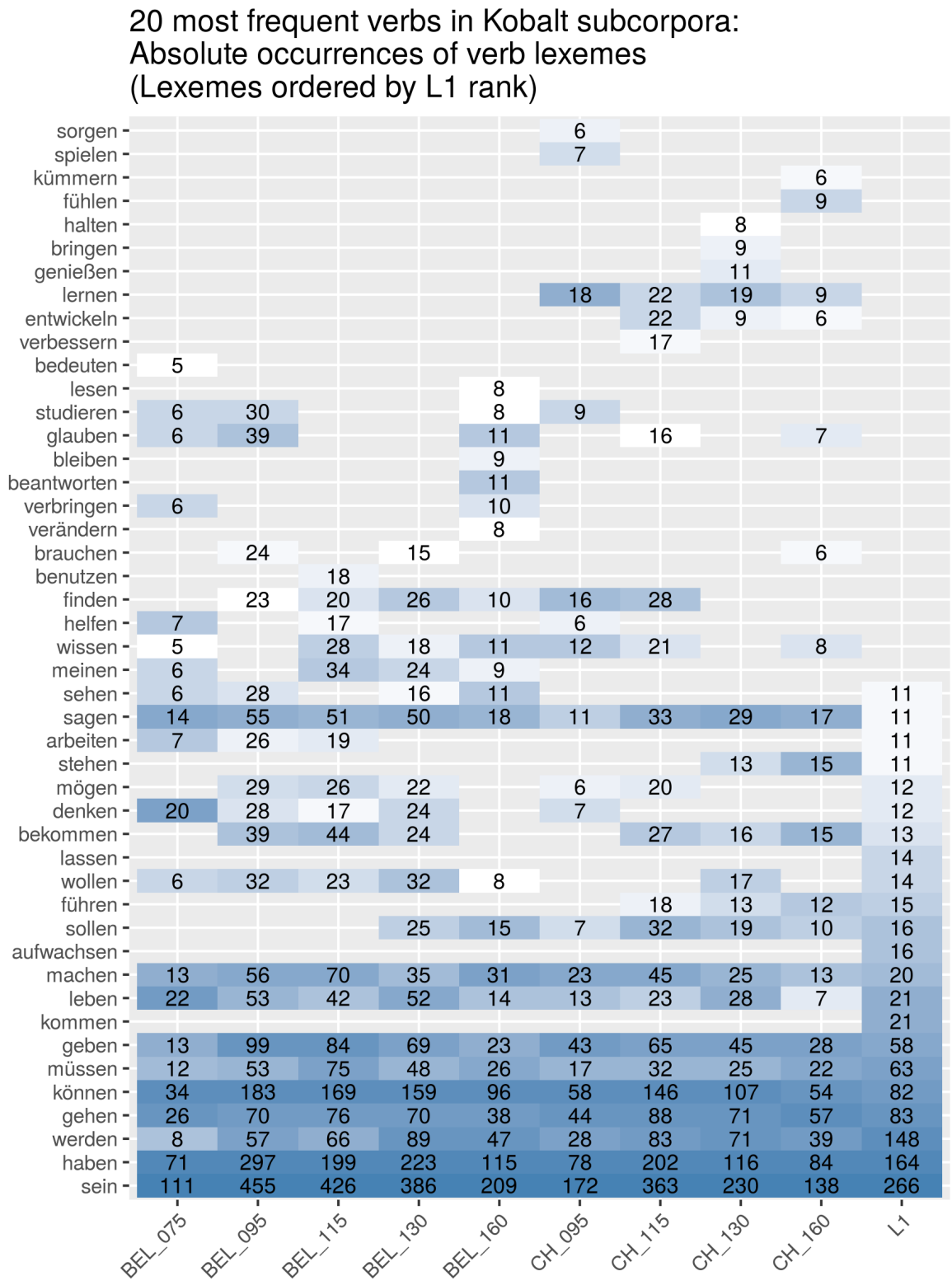


Figure 4.15.: Absolute frequencies of 20 most frequent verbs in Kobalt subcorpora. Colors indicate rank.

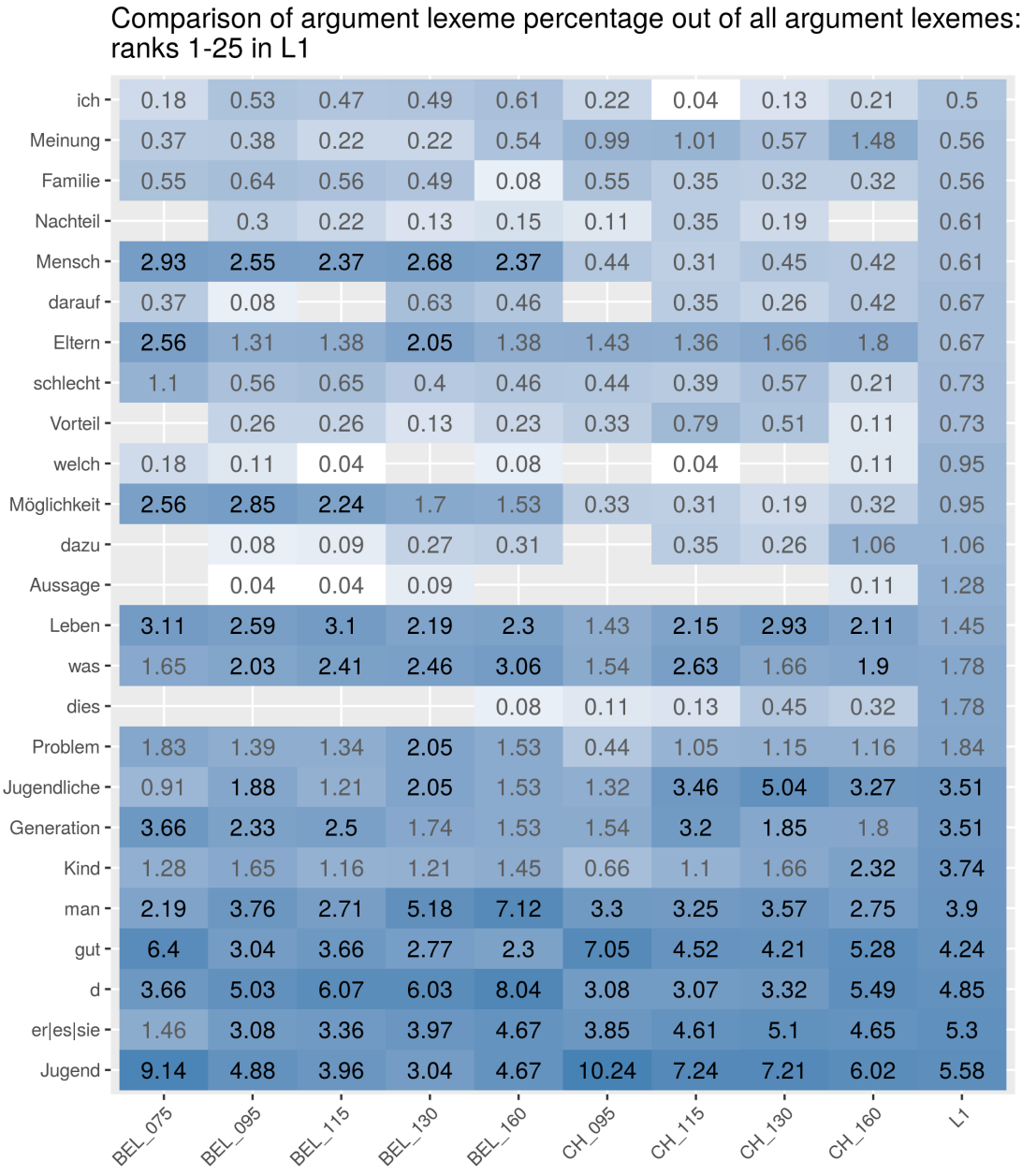


Figure 4.16.: Comparison of percentages of specified argument lexemes out of all argument lexemes, undivided by dependency slot, ordered by rank in L1. Ranks 1-25. For example, *Jugend* ('youth') makes up 5.58% of all arguments in L1 and is ranked first, while in CH-095, it makes up 10.24% of the arguments, and in BEL-130, it makes up only 3.04%.

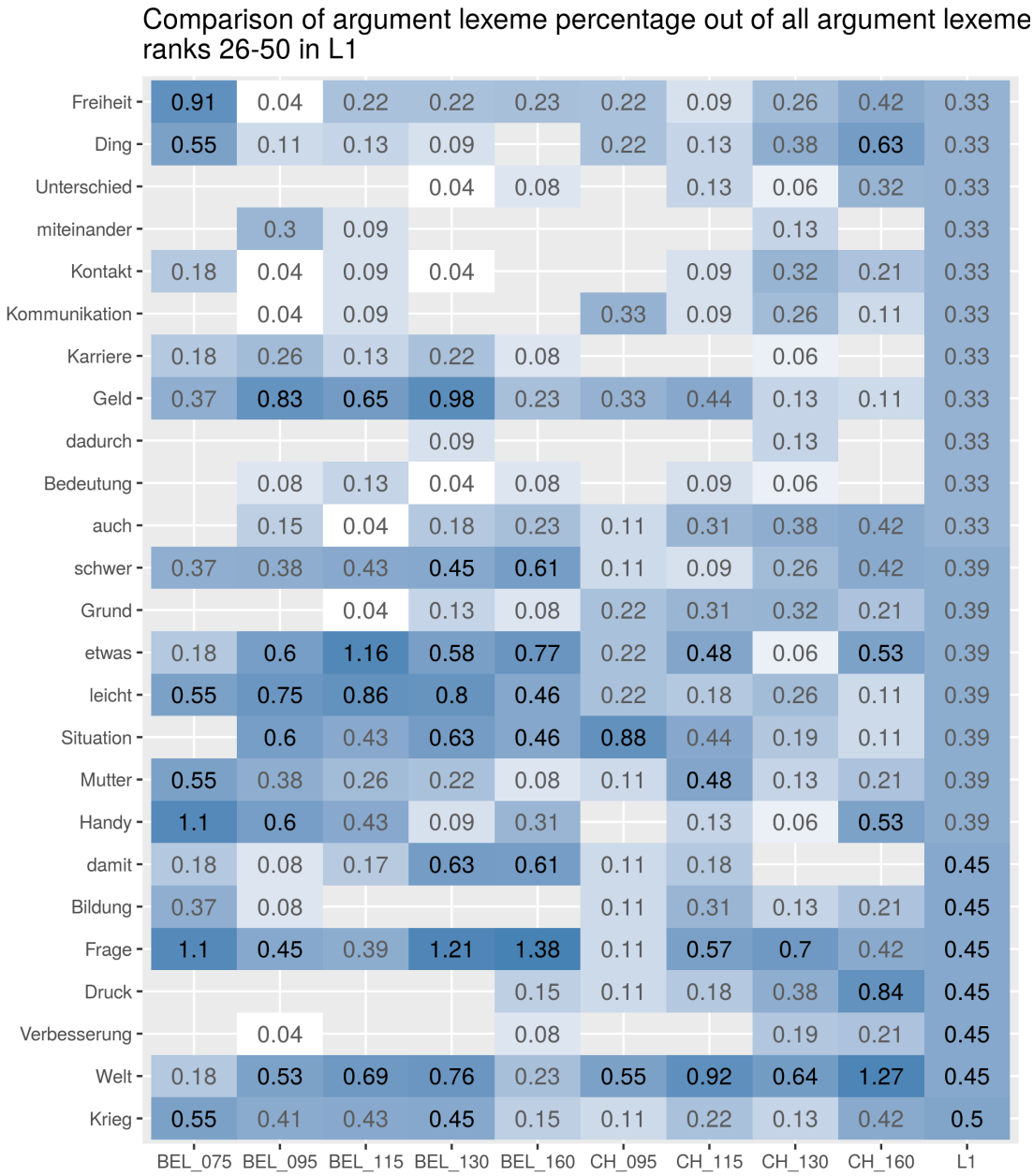


Figure 4.17.: Comparison of percentages of specified argument lexemes out of all argument lexemes, undivided by dependency slot, ordered by rank in L1. Ranks 26-50. For example, *Krieg* ('war') makes up 0.5% of all arguments in L1 and is ranked 26th, while it makes up only 0.11% of the arguments in CH-095. Some arguments appear less often in the learner corpora now. Two of those (*miteinander, dadurch*, 'with one another', 'through that/which, that's why') are functional. The others are divided by language groups: *Druck* ('pressure') and *Bildung* ('education') is not used in all BEL subcorpora, while *Karriere* ('career') is not commonly used in CH.

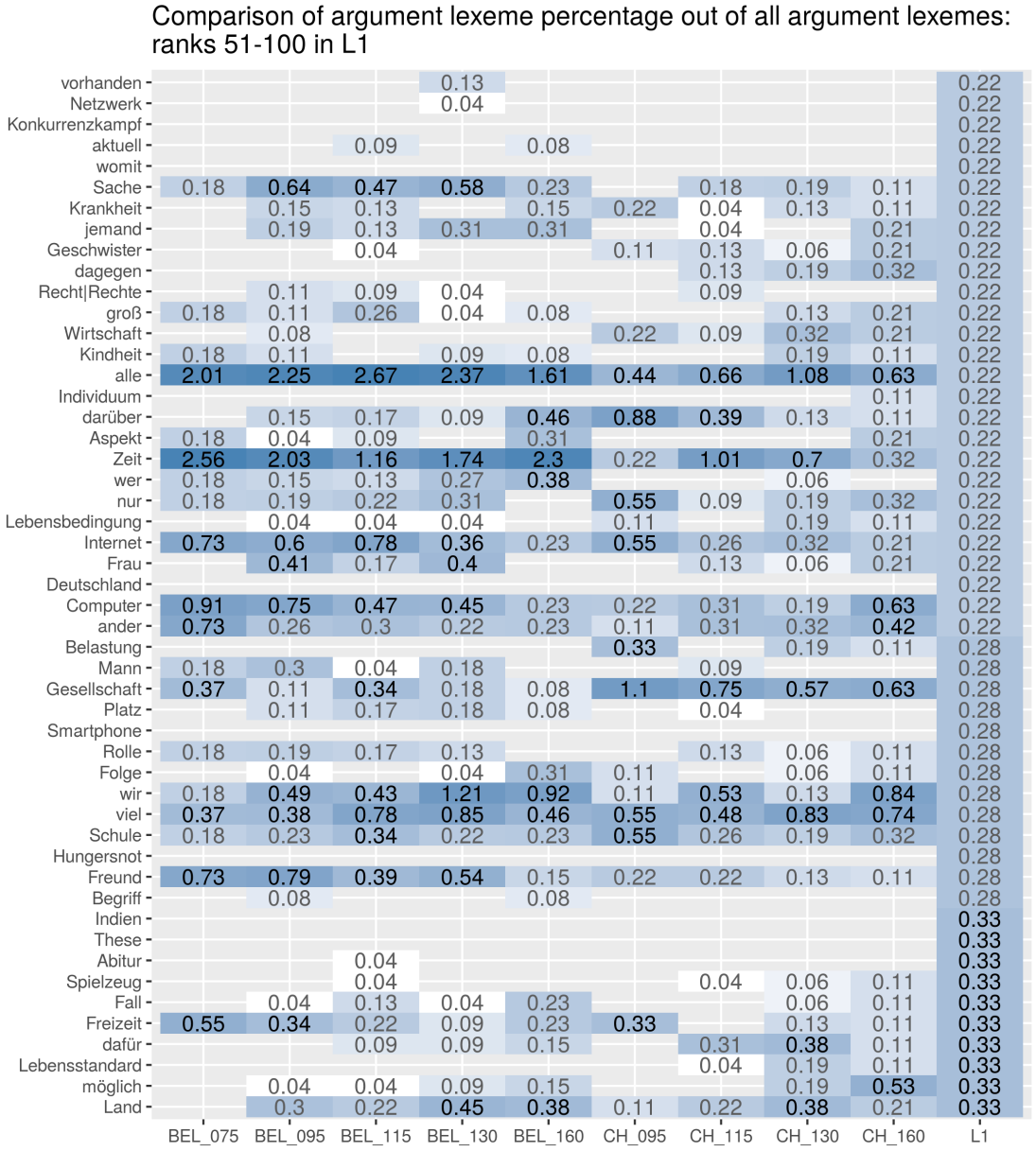


Figure 4.18.: Comparison of percentages of specified argument lexemes out of all argument lexemes, undivided by dependency slot, ordered by rank in L1. Ranks 51-100. For example, *Land* ('land, country') makes up 0.33% of all arguments in L1 and is ranked 51st, while it makes up 0.45% of the arguments in BEL-130.

4.2.4. Shared coselections

It has now been shown that there is a shared lexicon between subcorpora. However, this may not be reflected in many shared coselections. The combinatorial power of a coselection slot is the number of verb lexemes times the number of potential arguments. For example, with 304 unique lexemes in the OBJA slot in BEL-115 and 148 unique verb lexemes that occur with OBJA arguments, there are 44 992 potential verb + OBJA coselections (see figs. 4.1 and 4.2). This is counted from a usage-based perspective, i.e. the potential of only the vocabulary that is already accounted for in the corpus. Out of these 44 992 coselections only few can be drawn, namely the number of verbs that take accusative objects (726 in BEL-115).

Thus, the problem is as follows: If each combination had the same probability of $\frac{1}{44992}$, the probability of randomly drawing the same coselection twice would be $\frac{1}{44992^2}$, three times, $\frac{1}{44992^3}$, and so on. There are $\frac{(44992+726-1)!}{726!(44992-1)!}$ possible combinations that can be drawn from this set, which amounts to $1.3527 \cdot 10^{1617}$ (with replacement, coselections can occur 0–726 times, permutations are not counted separately). Even with a harsh limitation of the combinatorial power to 1% of the possible coselections (449, suggesting all of the other ones are semantically blocked – likely an overestimation), there are still $2.276 \cdot 10^{337}$ left. If furthermore one were to say that there are not 726 choices (as many as there are verbs selecting accusative objects), but only 72, because one assumes that 90% belong to highly frequent and highly coselectionally constrained verbs and are not chosen freely from the number of possible combinations – unique sets of 72 coselections out of 449 – is *still* at $3.352 \cdot 10^{89}$. For reference: The number of atoms in the universe is estimated to be somewhere between 10^{78} and 10^{82}. Lowering the number of draws further, suggesting that only 7 out of 726 verb + argument coselections underly free choice, *still* leaves $7.647 \cdot 10^{14}$ combinations of coselections out of 449. Against these staggering numbers, *any* unique combination of coselections is highly unlikely.

Of course, the *idiom principle* (Sinclair, 1991) states that coselections are not determined by the combinatorial power of the lexicon, but by convention. This would suggest that the same coselections appear across subcorpora, or at least across speakers in a subcorpus. It does not, however, provide a quantification of such limitations. How many shared coselections would it take for texts to be considered constrained vs. random in this combinatorial space? Tab. 4.4 shows the 12 coselections that occur in all subcorpora with their absolute frequencies.

- Six out of these are clearly coselectionally constrained, although not necessarily by the idiom principle, but by the prompt (all combinations of *Jugend, Jugendliche, Generation* ('youth', 'adolescents', 'generation'). *Eltern + sein* ('parents' + 'to be') is similarly topic-related, if not directly taken from the prompt);
- One is functional (*d + sein* 'this, that, which, who' + 'to be');
- Four can be considered coselectionally constrained as suggested by the idiom principle: *man + sagen* ('one' + 'to say') as in *one could say*; *Zeit + haben* ('time' + 'to have'); *(der) Meinung + sein* ('to be of the opinion'); *Problem + geben* ('problem' + (existential) 'to give' as in *there is a problem*).

Does this reflect high or low coselectional constraint? If considered separately by language group, there are 40 coselections that occur in all BEL subcorpora and 36 that occur in all CH subcorpora (tab. 4.5 and 4.6). Out of these, 27 and 23 respectively also occur in

arg	v	dep	CH-95	CH-115	CH-130	CH-160	L1	BEL-75	BEL-95	BEL-115	BEL-130	BEL-160
d	sein	SUBJ	8	17	10	9	9	4	34	60	40	24
Generation	haben	SUBJ	2	13	8	3	12	4	10	10	5	2
Generation	sein	SUBJ	2	17	1	1	4	9	22	10	7	3
Jugend	gehen	OBJD	25	57	43	29	37	17	35	31	33	18
Jugend	haben	SUBJ	11	15	10	5	13	11	27	20	13	12
Jugend	sein	SUBJ	18	40	18	3	2	8	20	15	7	10
Jugendliche	haben	SUBJ	4	15	15	2	14	2	11	6	8	4
man	sagen	SUBJ	4	7	12	2	5	2	13	5	7	3
Zeit	haben	OBJA	1	8	2	1	2	2	11	3	5	3
Eltern	sein	SUBJ	1	2	4	2	2	2	1	2	1	1
Meinung	sein	OBJG	1	1	1	5	4	1	3	1	1	3
Problem	geben	OBJA	2	4	2	1	1	2	4	6	8	1

Table 4.4.: The 12 coselections that occur in all ten subcorpora of Kobalt and absolute frequencies.

L1.[11] This could express a tendency to cluster by language, but it might as well be an effect from adding another subcorpus to the comparison, where with each additional subcorpus, the intersecting set shrinks. This is particularly likely since most of the coselections only occur once in most of the subcorpora.

Most of the coselections in these lists are of the high-frequency and semantically light verbs *haben, sein* or *geben* ('to have', 'to be', 'to give, to exist'). These would be well explained without referring to coselectional constraint, i.e. through the probable occurrence of a frequent verb and a frequent argument.

Are identical coselections across corpora then random, or are they coselectionally constrained? If the set was limited to 726 coselections from which one could choose, there would be exactly one combination in which each coselection occurred exactly once (a maximally productive model where all coselections are unique), but 725 if one occurred twice and all the other ones once, and 262 450 if one occurred three times. Which of the left-out coselections in the second and third case would make a set more or less coselectionally constrained compared to the first case? Or would all be equally coselectionally constrained, because it is only the re-occurrence of the first coselection that makes the constraint?

Arguing against this combinatorial power that re-occurring coselections are not chosen randomly is trivial: Randomness is obviously a poor baseline for coselection in language. But what would be a good baseline? An answer to this requires a *quantiative* model of the *idiom principle*, which does not exist to date. What can be said is that despite similarities in syntactic and lexical distributions, in absolute numbers, there are few coselections that are shared between subcorpora.

[11] Plots are included in the repository (`10.5281/zenodo.3584091`).

arg	v	dep	BEL-75	BEL-95	BEL-115	BEL-130	BEL-160
Computer	haben	OBJA	2	8	3	3	2
d	sein	SUBJ	4	34	60	40	24
d	bedeuten	SUBJ	1	2	3	1	1
Generation	haben	SUBJ	4	10	10	5	2
Generation	sein	SUBJ	9	22	10	7	3
Handy	haben	OBJA	4	11	4	1	4
Internet	haben	OBJA	1	6	1	4	1
Jugend	gehen	OBJD	17	35	31	33	18
Jugend	haben	SUBJ	11	27	20	13	12
Jugend	sein	SUBJ	8	20	15	7	10
Jugendliche	haben	SUBJ	2	11	6	8	4
Kind	haben	SUBJ	1	3	3	2	2
Kind	sein	SUBJ	1	5	2	1	1
Leben	sein	SUBJ	7	13	13	6	2
Leben	machen	OBJA	1	5	12	2	1
man	können	SUBJ	1	2	6	5	1
man	sagen	SUBJ	2	13	5	7	3
man	sprechen	SUBJ	1	1	1	1	1
Möglichkeit	haben	OBJA	11	54	38	21	12
Zeit	haben	OBJA	2	11	3	5	3
Zeit	sein	PRED	4	4	1	1	6
Zeit	sein	SUBJ	2	6	4	6	3
Zeit	verbringen	OBJA	2	10	5	2	7
Eltern	sein	SUBJ	2	1	2	1	1
Freiheit	haben	OBJA	2	1	3	2	2
Meinung	sein	OBJG	1	3	1	1	3
alle	haben	OBJA	2	5	2	5	2
alle	haben	SUBJ	1	4	1	2	5
alle	sein	SUBJ	4	7	16	10	3
Angst	haben	OBJA	1	8	6	5	3
Mensch	haben	SUBJ	4	10	8	7	4
Mensch	sein	SUBJ	3	12	7	7	3
Problem	geben	OBJA	2	4	6	8	1
Problem	haben	OBJA	6	11	4	11	3
Frage	sein	PRED	4	3	4	6	2
Freizeit	verbringen	OBJA	3	5	3	1	1
Leute	sein	SUBJ	4	6	5	3	1
Recht	haben	OBJA	1	4	1	5	1
Ausbildung	bekommen	OBJA	1	6	9	5	2
Buch	lesen	OBJA	3	5	6	4	3

Table 4.5.: All coselections that occur in all BEL subcorpora of Kobalt and absolute frequencies.

arg	v	dep	CH-95	CH-115	CH-130	CH-160
d	sein	SUBJ	8	17	10	9
d	heißen	SUBJ	1	4	5	4
Generation	gehen	OBJD	1	2	4	2
Generation	haben	SUBJ	2	13	8	3
Generation	sein	SUBJ	2	17	1	1
Jugend	gehen	OBJD	25	57	43	29
Jugend	haben	SUBJ	11	15	10	5
Jugend	sein	SUBJ	18	40	18	3
Jugend	werden	SUBJ	1	1	2	1
Jugend	führen	SUBJ	1	2	5	2
Jugendliche	haben	SUBJ	4	15	15	2
Jugendliche	sein	SUBJ	1	13	4	4
Jugendliche	machen	SUBJ	1	1	1	1
Kind	geben	OBJA	3	1	4	3
Leben	genießen	OBJA	1	5	1	1
Leben	führen	OBJA	3	7	11	6
man	haben	SUBJ	4	4	3	1
man	sein	SUBJ	1	2	2	1
man	sagen	SUBJ	4	7	12	2
man	machen	SUBJ	1	5	1	2
Schule	gehen	OBJP	3	3	1	3
Seite	haben	OBJA	2	4	2	1
Welt	sein	SUBJ	1	2	2	1
Zeit	haben	OBJA	1	8	2	1
Eltern	haben	SUBJ	2	2	2	2
Eltern	sein	SUBJ	1	2	4	2
Meinung	sein	OBJG	1	1	1	5
Problem	geben	OBJA	2	4	2	1
Wirtschaft	entwickeln	SUBJ	1	1	2	1
Chance	haben	OBJA	6	19	10	2
Grund	sein	PRED	1	2	3	1
Geschwister	haben	OBJA	1	2	1	1
Gesellschaft	entwickeln	SUBJ	2	3	2	1
Antwort	sein	SUBJ	1	2	3	2
Buch	lesen	OBJA	4	3	1	2
Nachricht	lesen	OBJA	1	1	1	1

Table 4.6.: All coselections that occur in all CH subcorpora of Kobalt and absolute frequencies.

4.3. Specialization and lexical association

It has been shown in section 4.1.2 that there are hundreds of unique coselections per slot, and in section 4.2.4, that only 12 of them occur in all subcorpora, and even within language groups, only 36 and 40 occur in all subcorpora. Out of those, most are coselections of the verbs *haben, geben,* or *sein*, and most occur only once or twice in most subcorpora. It seems then that a comparison of *the same* coselections across all corpora will not yield satisfactory results with respect to the research question.

However, coselectional constraint *within* subcorpora should still be quantifiable. This would mean to abstract from the item-based focus of a comparison of the same items across corpora. However, it would still be an item-based approach. While the research question is aimed at structural aspects without relying too explicitly on individual items, in a statistical operationalization, structure would always have to be derived from the items' individual behavior.[12]

In quantitative approaches, coselectional constraint or selectivity has often been modeled as high statistical dependency of two lexemes, for example by Evert (2005); Gries and Stefanowitsch (2004); Stefanowitsch and Gries (2005); Evert et al. (2017); Hilpert (2006); Gregory et al. (1999) among others. This bears a complication here, since with absolute numbers of co-occurrences in syntactic slots quickly declining towards less than ten in each subcorpus, large differences in the statistics may be the product of small differences in the text and in the linguistic model: An absolute frequency of eight vs. six is a decline by 25%, but cannot reasonably be considered as highly revealing regarding coselectional constraint. This is especially so because of the characteristic of burstiness that shapes language and text, a term describing the tendency of items to cluster in a certain window of a time series. In text, this refers to words re-occurring once they are introduced, for example for coherence or through priming, suggesting a high frequency within a small window that is not necessarily an expression of a high overall frequency (Sharoff, 2017; Hilpert, 2017, 60ff.).

A difference of three vs. six identical occurrences may relate to differences in coselectional constraint, or it might reflect a bursty development, i.e. one where two learners introduce the same coselection and then reuse it twice. Modeling this is possible through counting the number of documents a coselection appears in rather than the absolute frequency of occurrences, or by applying a reuse penalty on the count or excluding re-occurrences from a certain token window from the analysis. Yet none of those would solve the problem of all items belonging to the same order of magnitude.

But even in the largest corpora, due to the long-tailed Zipf-distribution, half of the words are hapaxes and therefore will also only co-occur with a verb only once, and most of the other half twice, meaning that the general problem of comparability of magnitudes exists in larger data as well.[13]

[12] Even if they were subsumed in categories, the categories would still be compared individually.

[13] Krenn (2000, 369) compares the performance of some lexical association measures in data that includes low frequency data vs. data that does not and concludes that

> "with decreasing co-occurrence threshold and increasing proportion of low frequency data among the collocation candidates, precision of the statistical measures deteriorates. For an increase of identification accuracy, the following two strategies may be pursued: 1. consider only word combinations with high co-occurrence frequency, then apply statistical models (...); 2. on the one hand apply [specific measures] on the complete data set and select the highest ranked word combinations, on the other hand select the highest ranked word combinations according to co-occurrence frequency. Combine the two sets. While 1. leads to a stronger increase in precision, a much higher number of collocations is identified

4.3.1. Statistical measures vs. the linguistic model

Where idiomaticity or coselectional constraint has been measured in a model of statistical (in-)dependence, this has usually been done with lexical association measures derived from the relative frequencies of co-occurrence of two lexemes in a corpus. This has been criticized in a number of ways, most recently in a summary by Gries (2019), in which he names the problem of the conflation of frequency and effect size, the conflation of distributions, the symmetry of most measures (more on that below), and the underdispersion of co-occurrences in a corpus, meaning that for some co-occurrences, the frequencies stem from only a small part of the corpus but are treated as if they were representative of the whole corpus. This final problem goes beyond what Gries describes though:

The mathematical model behind the concept of lexical association is that a lexeme w_1 has a (stable) probability $P(w_1)$ to occur in language L while a lexeme w_2 has a (stable) probability $P(w_2)$ and that the conditional probability $P(w_1|w_2) \neq P(w_1)$ and $P(w_2|w_1) \neq P(w_2)$ for collocations. The larger the difference $P(w_1|w_2)$ - $P(w_1)$ or $P(w_2|w_1)$ - $P(w_2)$ (and transformations thereof), the higher the collocational strength between two lexical items.

This is not actually a very good representation of language, because it builds on assumptions of randomness and independence, which are trivially wrong in language (Kilgarriff, 2005); and on the assumption that a sufficiently large corpus is an adequate representation of a system that is overall well described in terms of probabilities, namely a stationary ergodic system (Manning and Schütze, 1999, 76). Stationarity refers to the property of a system to exhibit the same means and variance across time, while ergodicity describes the property of a system to return equal statistics for any of a defined number of processes over time, i.e. in the words of Manning and Schütze (1999, 76), that a system "cannot get into different substates that it will not escape from". In stationary ergodic systems, probabilities are stable and independent of the point of measurement, and limits of relative frequencies for all variables are reached reliably, no matter which part or time of the system they are taken from: Regardless of whether I count dice rolls 1-10000 or 5000-15000, relative frequencies should always approximate expected values. A non-ergodic system is one where the outcome of one experiment defines the space of potential outcome of others, i.e. in which the initialization defines a path that may not allow for the overall system to reach expected values any longer. For example, if I count dice rolls 1-5000 as '1', because the first roll was a 1, 2, or 3, and then randomly choose a new number, the system is not ergodic. Language, with priming, burstiness, interdependence of different linguistic levels, phenomena of alignment or convergence between speakers, and language change, is likely not to be an ergodic system (see Dębowski (2018) for a mathematical perspective).

Language cannot be stationary for open lexeme classes, since with the appearance of any new lexeme, the relative frequencies of *all* lexemes change, therefore denying them the ability to converge to predictable values. But even without accounting for productivity and wider-scale language change, relative frequencies of most words, especially where dynamic topics are treated like in newspapers or in a newswire[14] can in fact not be expected to *ever*

employing 2."

Both of these are workaround solutions for the deeper problem of modeling a long-tailed distribution while still assuming that all items have an equal potential force of attraction on other items. The model is contradictory, because it builds on frequency while also allowing for frequency to not be decisive.

[14]This is the odd example used by Manning and Schütze (1999, 76) as approximating stationarity in the period of one year

approximate limits in a changing world. This is obvious for some idiosyncratic words like the *Y2K-bug* that quickly became redundant after January, 1st (or perhaps 2nd or 3rd) of the year 2000. But even in less obvious cases, many content words will change frequency in a changing world and leave unclarity about expected values.[15]

This is relevant, because measures built on conditional probabilities presume that the sample used is sufficiently large for relative frequencies to approximate their limits, i.e. probabilities. The mathematical foundation of this is the central limit theorem. Stating that in a chance experiment approximating infinity, relative frequencies of all events will approximate their limits regardless of the underlying distribution, and thus variables will reach expected values, it secures the connection between frequent, past observations and valid expectations for future outcome. If expected values keep changing, they cannot be reached through convergence, which also means that the past outcome of experiments is not predictive of the future: This is also why most scholars would not overall describe history as a stochastic process – anything that evolves is not stationary, and likely not ergodic. If the expected frequencies of words or coselections change – as they do in language over time – or if different parts of the system have different expected values – as is the case for lexical material in interaction with register and topic – accordance with the central limit theorem cannot be safely assumed. With this, all concepts related to statistics (probability, expected value, significance, effect strength, variance) face problems of definition, and this is unsupportive of the validity of a statistical analysis.

Importantly, this is not due to the properties of the underlying distribution – a distribution can be bi- or multimodal, and still yield stable probabilities, namely the average of all modes. The question is whether an underlying *stochastic* system can be reasonably assumed, i.e. a system which consistently yields the same relative frequencies over time (even if time is narrowly defined, this is not clear).

Consider the following thought experiment: Assuming that a large corpus is the result of a stationary and ergodic system, this means that it reflects the necessary result of the underlying stochastic system that is language (or a sublanguage, like a register). This then is to say that, given the conditions that the stochastic system reacts to, it *could not have produced meaningfully divergent relative frequencies*. This is to imply a a bizarre determinism: Since a stochastic system is deterministic in outcome (=expected value *will* be approximated), people could not have written texts different from the ones they did beyond a rearrangement of elements and mild fluctuation in word usage. This model would also require a placeholder for productivity: The stochastic system has always held a separate set of probabilities for words which did not exist yet, and those were determined to be a certain amount from the beginning.

It should be mentioned that this is not a matter of appreciation of statistical methods or one of taste: If lexemes do not have probabilities, they cannot be entered into a mathematical model *requiring* probabilities regardless of methodological preference. In the same way that it would be difficult to define a temperature or a velocity for a written word – which means that it cannot be compared with the volocity or temperature of other entities – if an entity does not have a feature of a describable probability, mathematical models hinging on probability are *undefined*. Thus, the underlying question concerns the

[15] In fact, Baayen (2001, 36f., my emphasis) makes this point implicitly in writing that: "[t]he vocabulary richness of a given text is a problematic concept which suffers from the dependence of the majority of procedures on text length N. The longer the text, the (relatively) smaller the increase of different words (V) in it. Hence if a text is "sufficiently" long, the rate of change of the majority of type-token indices, dV/dN, must converge to zero. *If they converge to infinity, they are in principle wrong,* though measurement in finite texts is possible".

ontology of language and linguistic study: Is there, *at all*, a stochastic system in the first place – do lexical items have probabilities, or do they fluctuate a lot? And, if such a stochastic system exists, is *any* corpus a good representation, no matter how fine-grained or narrowly defined it is – i.e. is the system stable enough, and if so, over which time and space (register), so that it may be sampled from? And finally: Is corpus compilation a good sampling function of this system?

Indirect evidence against the presence of stationarity comes from applications of natural language processing (NLP) where different levels of accuracy are reached in various in-domain samples, i.e. algorithms perform unequally well on different splits *of the same kind of data that they were trained on* (Barrett et al., 2019). Similarly, Piantadosi (2014) shows that, even within the same corpus, even if it is large, relative word frequencies do not appear converge.

Let it be said that these observations are limited to the scope of lexical distributions for the purposes of this thesis. Syntactic and morphosyntactic probabilities are different for two reasons: Firstly, they are more stable over time and space, and they seem to converge relatively quickly (as has been shown in the POS distribution earlier in this chapter). This is not to say that with phenomena like grammaticalization, constructional extension, morphosyntactic assimilation in language contact contexts, etc., stationarity and ergodicity are a cross-systemic, unchanging feature. However, secondly, more abstract linguistic features compare categories rather than exemplars. The development of a whole new category (such as loss or gain of a case system, for example) *does* notably shift a linguistic system in complex ways, and could perhaps be modeled to redistribute probabilities. The productive use of a new word, on the other hand, does not nearly have the same effect, but mathematically those two are treated as equal in statistical approaches.

It is possible that there exist definable subsystems in language that are indeed both stationary and ergodic, even in terms of lexical distributions (such as highly register-, topic- and/or historically narrow corpora) or subsystems that oscillate between two or more ergodic states, but it seems unlikely that large corpora spanning many decades, registers, and topics are such systems. This aspect is not yet widely discussed in linguistics, but a discussion seems both necessary and worthwhile, at least where inferential text statistics are used in theory-building arguments.[16]

In the literature, the failure of the randomness assumption is frequently mentioned in passing, but its repercussions on the linguistic model are rarely discussed. A few notable exceptions are Kilgarriff (2005), Schmid and Küchenhoff (2013), and more recently, Koplenig (2017) in arguing against p-values for this very reason: Combinations which are linguistically unlikely or next to impossible (like *lesen+Freizeit*, 'read+spare time') are still expected to occur in the statistical model, and the part of the distribution unfilled by those is by definition redistributed to other items. This yields overly significant results and artificially raised effect strengths across the sample for basically all infrequent words, which in a Zipf-distribution are most words (Baayen, 2001, see chapter 5, 5.1.2 and 5.2 in particular for a mathematical description). In other words, with lexical distributions being the way they are, any *specific* coselections of infrequent words are unexpected by

[16] Ergodicity and time- and space-sensitivity, including historicity, and its repercussions on quantiative analysis and prediction are actually frequently discussed in other fields working with dynamic systems, the more obvious ones such as mathematics and physics (quantum physics especially), but also in economics (Durlauf, 1993; Pålsson Syll, 2012; Verbrugge, 2006; Martin and Sunley, 2012; Wallace, 2013, and apparently many others) neuroscience (Medaglia et al., 2011; Franco et al., 2007; Papo, 2013), and social and developmental theory (Molenaar, 2008; Lerner, 2012).

chance.

Also, since the more items exist in a random model, the less likely they are to co-occur (to find the specific other item in a larger set is less likely than a smaller one), adding word pairs are already present as coselections or of which both words are new raises association values for *all* coselections, both the old ones and the new ones. In other words, if I fill my corpus with unrelated material, all previous coselections will become more unlikely in total and thus will be more highly associated. If I fill it with the same material that is already present, I will also raise the strength of association of all items.

Mathematically, it is necessary to assume independence and randomness, because the degree of dependence of two items can only be shown against the background of a potential independence – but linguistically, the effects are odd.

The epistemology of lexical association measures, where they are used for linguistic theory building, is hence not very robust. They are, however, used widely in heuristics for the identification of collocation candidates in lexicography (Pecina, 2010; Petrović et al., 2010; El Maarouf et al., 2014; Evert, 2008); to correlate corpus frequency with behavior in psycholinguistic experiments, such as shorter reading or retrieval times for more strongly associated coselections, Wiechmann (2008); Camblin et al. (2007); and for applications in computational linguistics and text mining, such as topic modeling or text classification (Dias et al., 2000; Orliac and Dillinger, 2003; Lau et al., 2013). For these tasks, it is not the precise numbers or even their ranks that are of high interest, but the relations of those numbers to one another.[17] This does not rid the analysis of the mathematical and epistemological issues discussed, but its claim is less epistemological and its purpose limited to a specific application, namely as a relatively coarse filter. With these concerns, it appears that an analysis of lexical association would not provide an excellent ground for a quantification of the coselectional constraint in a subcorpus *in toto*. But it may provide insight into classes of coselections and differences and their distributions in subcorpora. This in turn may synthesize into a clearer understanding of the overall development.

4.3.2. Lexical association in Kobalt

Measuring forces of attraction between words statistically marks an item-based approach in that association is measured individually for each coselected pair of items. But all statistical lexical association measures are derived from the distribution of not only the items in question, but all other items in a corpus, meaning that they are not in fact simply item-based as in 'independent of other items'. Rather, frequencies of co-occurrence, of individual occurrence of each item, and of non-occurrence of the items in question are entered into contingency tables, n x n matrices where n denotes the number of items of which the association is measured (two in this analysis). See tab. 4.7 and tab. 4.8 for an example.

The long list of lexical association measures that exist in the literature based on this model cannot be discussed here. For a more recent comparison of some frequently used measures, see Evert et al. (2017). Generally, measures differ by task-based performance, most commonly in collocation extraction in applied computational linguistics and lexicography. As Bouma (2009, 39) notes, "the collocation literature has shown that the effectiveness of a measure is strongly related with the task", meaning there is no single

[17] Ranks 1 and 100 may be apart by 100 occurrences in absolute numbers, and by 99 ranks, but importantly, one is 100 times as frequent as the other. This could also be expressed in ranks 200 and 500 or in absolute frequencies of 125 and 1250.

a \ b	b	¬b	sum
a	a ∧ b	a ¬ b	total a
¬ a	b ¬ a	¬ a ¬ b	total ¬ a
sum	total b	total ¬ b	total structures

Table 4.7.: Contingency table for factor combinations such as word co-occurrences

haben \ Problem	Problem	¬Problem	sum
haben	5	93	98
¬ haben	4	452	456
sum	9	545	554

Table 4.8.: Contingency table for the coselection of *haben + Problem* ('to have + problem') in Kobalt L1

measure that is equally well suited for large and small corpora, rare and frequent words and different combinations thereof. It should also be noted that all lexical association measures that are based on conditional probabilities are derived from the same data in one way or another. This means they are in most cases transformations of one another, so while they may treat numbers of different magnitudes in different ways, none of them have access to extra information layers that would change results in a more profound way.

In fact, the only measure that has recently been suggested that works unlike this – although it still depends on the same table, but in a slightly different way – is the one chosen here: ΔP, which is a two-sided or directional measure with two values in the range of $[-1, 1]$ for each coselected pair. ΔP(verb|argument) is the conditional probability of the verb to occur with the argument minus the conditional probability of the verb to occur without it (= with another argument); and ΔP(argument|verb) is the conditional probability of the argument to occur with the verb minus the conditional probability of the argument to occur with another verb:

$$\Delta P(verb|arg) = P(verb|arg) - P(verb|\neg arg)$$
$$\Delta P(arg|verb) = P(arg|verb) - P(arg|\neg verb)$$

In the contingency table, this translates in the following way: If the upper left field is labeled a, the upper right is b, the lower left is c, and the lower right is d,

$$\Delta P(verb|arg) = \frac{a}{a+c} - \frac{b}{b+d}$$
$$\Delta P(arg|verb) = \frac{a}{a+b} - \frac{c}{c+d}$$

The measure was introduced into the linguistic discussion first by Ellis (2006a), but discussed for text statistics in depth in Gries (2013). Gries suggests it as a valuable alternative to most unified measures that do not make a distinction between the force of attraction one item has on another, which may differ from the reverse direction as in his example of 'of' to 'course' (P(of|course) > P(course|of), Gries (2013, 144)), where *course* attracts

of strongly, unlike the much more frequent *of*, that does not equally attract *course*. In other words, if I see 'course', I can reliably guess that 'of' will occur, but not vice versa. This is particularly interesting in a lexicosyntactic coselection research question. While collocation is typically analyzed as two words occurring together or retrieved at the same time, another definition of coselection would be to view one word as being semantically selected and the other as coselected, basically entailed by the other, like a +1 to a party. This process would not have to be, and likely would not be, symmetric for both items. A directional measure would then be helpful in determining the forces of both.

ΔP is however also a measure that is a bit odd in terms of stochastic theory, because it subtracts two probabilities that are based on a different totality (for example, the total of the verb slot might be 700 and the total of the argument slot 2000). To give a non-linguistic example, in

$$\Delta P(I\,win\,the\,lottery|It\,starts\,to\,rain) = P(I\,win\,the\,lottery|It\,starts\,to\,rain) - P(I\,win\,the\,lottery|It\,does\,not\,start\,to\,rain)$$

the first figure is the conditional probability of my winning the lottery while it also starts raining, which is the number of times I win the lottery and it starts to rain divided by the number of rainy times; while the second is the number of times I win the lottery and it starts to rain divided by the number of unrainy times. Subtracting one from the other yields a figure between [-1,1], which may express some force of attraction or repulsion between the two events, but not a probability, and not a transformation of a probability either. Probabilities that do not inhabit the same stochastic space cannot be subtracted without leaving said space (because probabilities inhabit a space in [0,1] and the ΔP space is [-1,1]) and subtraction is not clearly defined for probabilities relating to different stochastic spaces. What is, in layman's terms, *the probability that I win the lottery and it starts to rain* minus *the probability that I win the lottery and it does not start to rain*? Probabilities can be translated to idealized materializations of a case vs. possible cases: In x out of y cases, z happens. but with ΔP, the conceptual meaning seems less clear. This is also because a ΔP of 0.25 can stem from a number of different linguistic cases 0.75-0.5 (if this argument occurs, three out of four times, this verb will also occur, but if another argument occurs, in half of the cases, the verb will still be there – it is just a very frequent verb) or a 1 - 0.75 or a 0.26 - 0.01 (given this argument, in roughly one quarter of the cases, the verb will occur, but if this argument is absent, only 1 in 100 times will this verb occur – it is apparently an infrequent verb outside of this argument and perhaps altogether).

Empirically, ΔP does seem to recognize collocation candidates, but with less accuracy than some other measures in Evert et al. (2017).[18] The main advantage of this measure is that being two-sided, it can be seen as offering a higher resolution, because each coselection can be assessed on two dimensions. This might be revealing in terms of which coselections are of interest for further analysis, and it is also, as shall be seen, a convenient way of visualizing coselections in what appears to be a linguistically meaningful way. The results

[18] It is possible though that ΔP recognizes (true) collocations that just happen to deviate from the gold standard which in this study was "a fixed set of known collocations" obtained from two collocation dictionaries (Evert et al., 2017, 534). The definition of collocation in those dictionaries may not include certain aspects of language in use, either through (statistically) false positives in the dictionary, where a syntactic or semantic, but not statistical collocation is considered one in the dictionary – such as the frequently mentioned example of 'to kick the bucket' that occurs very rarely in actual use; or through (statistically) false negatives in the dictionary, where statistically frequently co-occurring words do not appear in the dictionary but do have relevance for language in use.

will show that intuitively, ΔP seems to provide an estimation of how well $word_a$ can be predicted from knowing $word_b$ without assuming a symmetrical relationship like in pointwise *mutual* information (PMI). Whether this is a linguistically or mathematically valid intuition remains an open question at this point.

ΔP values were computed for all verbs and their accusative object, subject, prepositional object, and predicate slot fillers separately. A more fine-grained distinction into, for example, ditransitive or lexically specified constructions would be reduced to individual cases in this data, making a quantitative assessment virtually impossible.

4.3.2.1. ΔP(OBJA)

Results indicate that ΔP in Kobalt captures some relatively trivial[19] and some non-trivial coselectional aspects in the OBJA slot as can be seen in plots 4.19–4.24. These report coselections with a frequency ≥ 3 and ≥ 5.[20] The ΔP space of [-1,1] is not used entirely. Rather, only the upper right quadrant is filled in data of this size, suggesting a floor effect (repulsion cannot be measured, because all coselections are expected to occur rarely and cannot occur less than zero times).

- Most predictive in L1 are the reflexive verbs *beschäftigen, verändern* ('to engage', 'to change') predicting the reflexive pronoun *sich* as an OBJA and the OBJA *Problem* ('problem') predicting the verb *haben*. They are not representative of the highest ΔP values, but they do occur five times or more in the L1 corpus (see fig. 4.19).

- If all V+OBJA coselections occurring more than twice are considered (fig. 4.20), ΔP performs well as a categorizer in this dataset, but its scalar nature is barely expressed: It distinguishes well between arguments that are predicted by the verbs, reflexive verbs mostly, in the upper left quadrant, and verbs that are predicted by the argument, some of which are support verb constructions (*Kontakt pflegen* 'to maintain, to cultivate contact(s)', *eine Rolle spielen* 'to play a role', *eine Verbesserung darstellen* 'to mark an improvement', *Karriere machen* 'to make a career'). The reflexive verbs are not entirely trivial in German, because reflexive verbs can often also be used transitively as well with very similar semantics.[21] This means it would be possible for *verändern* ('to change') or *beschäftigen* ('to employ, to keep busy') to prefer other lexemes in its OBJA slot, but the reflexive use is apparently the most frequent.

- Beyond these two lexicosyntactically predefined groups, predictable coselections are mostly situated closer to zero on the x-axis, suggesting that a verb can be better predicted by the argument than vice versa (except for *Möglichkeit bieten* 'to offer an

[19] Much like in Gries (2013) where some of the strongest examples are *in vitro, not withstanding,* and *upside down* this illustrates an aspect that is also discussed in the wider digital humanities and computational literary studies: "In a nutshell the problem with computational literary analysis as it stands is that what is robust is obvious (in the empirical sense) and what is not obvious is not robust, a situation not easily overcome given the nature of literary data and the nature of statistical inquiry" (Da, 2019, 601). One might argue that this is a post-hoc observation though, and that results could not have been predicted from the theory. This is a strong objection that requires deeper methodological discussion in corpus-based lexicosyntax, because most studies do not clearly define their scope of evidence (exploratory vs. confirmatory, for example).

[20] All thresholds are chosen arbitrarily, but the lowest number that could possibly suggest a coselectional constraint is 2, but coselections of frequency 2 are too many to fit legibly into a plot.

[21] This is related to the concepts of unergativity/unaccusativity, see section 5.2.

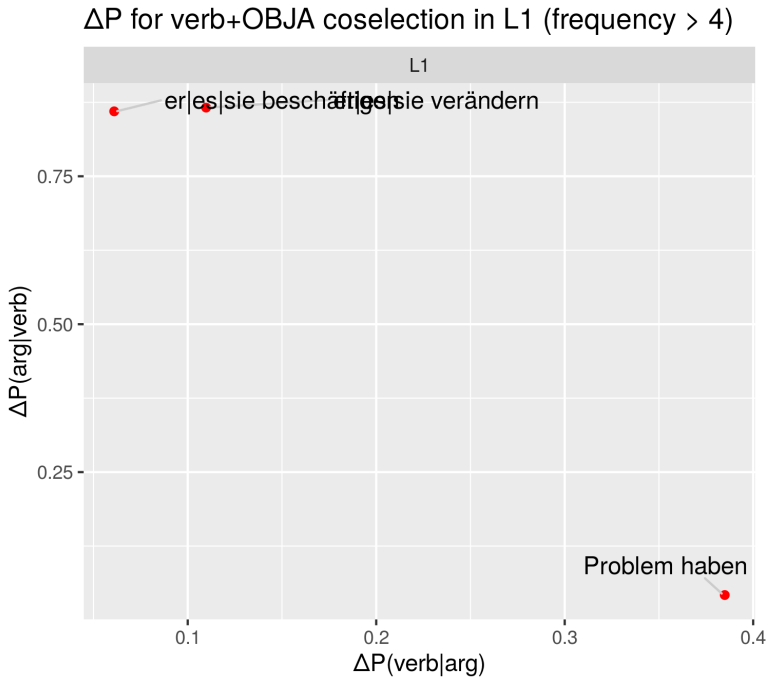


Figure 4.19.: ΔP for V+OBJA coselection in L1, frequency ≥ 5

opportunity', *Aussage treffen, diskutieren* 'to make, to discuss a statement'). Those coselections are all collocates of *haben* or *geben* ('to have', 'to give, to exist'), which is likely a statistical artifact of the high frequency of those verbs: If two verbs dominate the stochastic space, any argument will predict one of those reliably.

In BEL-learners, however, the picture looks different:

- There are many more identical coselections with a frequency ≥ 5, even for BEL-115 and BEL-130, which are both about 15% larger than L1 in tokens, but have 14 and 15 coselections with a frequency ≥ 5 respectively (compared to three in L1).
- Coselections are clearly of a different kind. While the coselection of *Problem+haben* ('to have+problem') that occured 5 times or more in L1 does appear in all BEL subcorpora except the last one, only two out of the five subcorpora show two reflexive verbs in the upper left quadrant.
- There is a high overlap in coselections between subcorpora, but all are quite generic. Only *Rolle spielen* ('to play a role'), *Zeit verbringen* ('to spend time'), *Angst haben* ('to be scared') and *Recht haben* ('to be right'), that occur in several subcorpora, may be considered coselectionally restricted in a collocational or idiomatic sense. Coselectional constraint has not been defined as a phenomenon of only clearly lexicalized cases in this thesis, but rather on the contrary was meant to include not obvious constraints. But it does show that the coselections of BEL-learners and L1-writers are qualitatively different.
- One interesting case is that of *Ausbildung bekommen* ('get an education'), which is unidiomatic in German (the idiomatic expression is *eine Ausbildung machen* and refers to vocational, professional or perhaps artistic training rather than, for example,

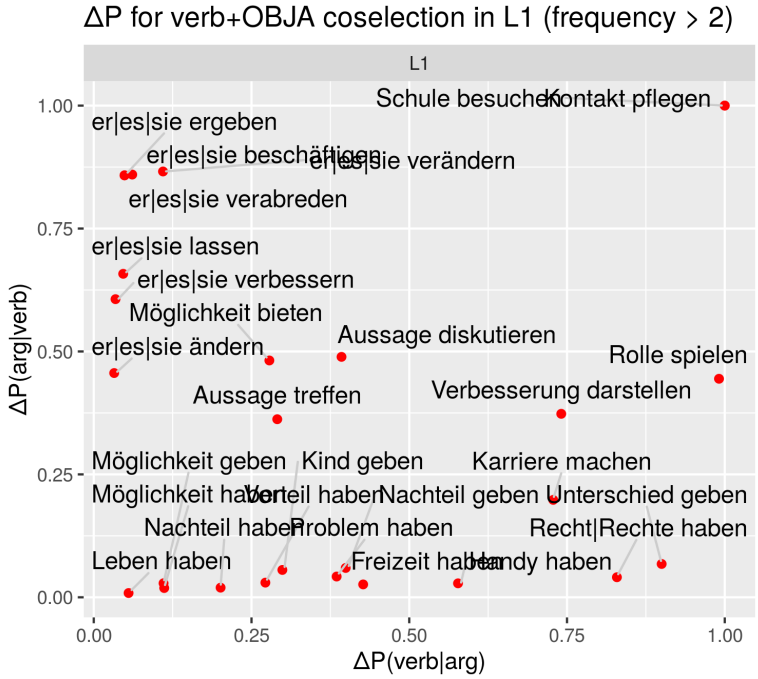


Figure 4.20.: ΔP for V+OBJA coselection in L1, frequency ≥ 3

university studies), but frequent and idiomatic in Belarusian (*атрымаць адукацыю*, 'atrymac adukacyu') and Russian (*получить образование*, 'polučit' obrazovanie'), which is the second official language in Belarus. The recurrent use in Kobalt is a likely case of L1-transfer, and also a lexical teddy bear that is overused even by upper intermediate learners (BEL-130 relates roughly to a B2.2 level as defined by CEFR, see section 3.2.1 for a discussion).

Regarding a potential u-shaped development, it is interesting to observe that BEL-75 and BEL-160 both have many fewer identical coselections, and not proportionally to the difference in size. This however would mark an invertion of the u-shape as it was discussed: Coselections seem more random and more generic in the intermediate corpora, but they are also many more. While this seems contradictory, it might be an effect of the predicted randomization or breaking up of fixed structures, where more frequent verbs (*haben* 'to have', *geben* 'to give, to exist', *bekommen* 'to get', *machen* 'to do, to make') instead of going with their previous fixed arguments now randomly select from frequent arguments (*Möglichkeit* 'opportunity, chance', *alle* 'all, everything, everyone', *d* 'the, this, that, which', *Problem* 'problem', was, *Geld* 'money').

Adding coselections of lower frequency to the plot (fig. 4.22) mainly shows the proliferation of identical coselection in the intermediate corpora (which are also larger than the marginal ones, though). However, for the BEL-130 and BEL-160 corpora, the picture does begin to resemble the one in L1 (fig. 4.20). In the upper left quadrant, there are more reflexive verbs, although they are scattered over a larger area, and in the lower right quadrant, some clearly noun-dominated coselections appear, like *Entscheidung treffen, Gedanke(n) machen, Zugang haben* ('to take a decision', 'to think, to worry (about)', 'to have access'). However, even in BEL-160, there are still a number of coselections that appear to stem from a more random distribution, like *d machen, was machen, etwas machen*

('to do this, that, something', 'what, which (they) do').

In CH-learners, co-occurrences of frequency ≥ 5 are less frequent than in BEL. Yet still, there are more than in L1 in both of the intermediate two subcorpora, now even despite CH-130 being roughly 15% smaller than L1. Coselections also seem to be of a more similar kind compared to L1:

- In CH-115, CH-130 and CH-160, reflexive verbs are well predicted through ΔP, and the support verb construction *Leben führen* occurs in three of the subcorpora.
- Some randomness is visible in the first three subcorpora too, though, namely *Leute geben* as in *Es gibt Leute* ('there are people'), *Zeit, Problem, Chance haben* ('to have time, a problem, an opportunity'), and *was machen* ('to do something' 'which, what (they) do' (interrogative, relative)).
- Considering coselections that occur three times and more, unlike in BEL, a number of coselections that are of a more lexicalized kind in L1 appear even in CH-115: *Wert legen (auf)* ('to put value on'), *Probleme lösen* ('to solve problems'), *Antwort geben* ('to give (an) answer'). However here, although in CH this is not the largest corpus, the same happens as in BEL-95, namely that a number of coselections appear that seem more random collocates of mostly *haben* and *geben*. CH overall appears again as lying between L1 and BEL.

One explanation for the earlier occurrence of L1-like coselections (among others) in CH might be typological, where Mandarin lacks a complex verb morphology and instead relies on V+NP complexes to convey complex semantics, and in a teaching effect, since rote learning and phrase lexicalization is more encouraged as a learning technique than is commonly the case in European schools and universities. This will further be discussed in chapter 7.

In summary, ΔP does work as a filter and a categorizer even for small numbers in OBJA. It shows a randomization, diversification, and a specialization in the use of lexicosyntactic coselection in OBJA slots, even with the low numbers of co-occurrence in this corpus.

It however also shows that a reliance on individual items, even on individual verbs, is not a great quantification for coselectional constraint even in individual subcorpora: The only verbs that occur frequently are *haben* and *geben* ('to have', 'to give, to exist'), and those are of such a generic kind that they do not provide stable ground for an argument around coselectional constraint. Thus, despite yielding interesting qualitative results, it does not provide a clear perspective for an operationalization of the research question.

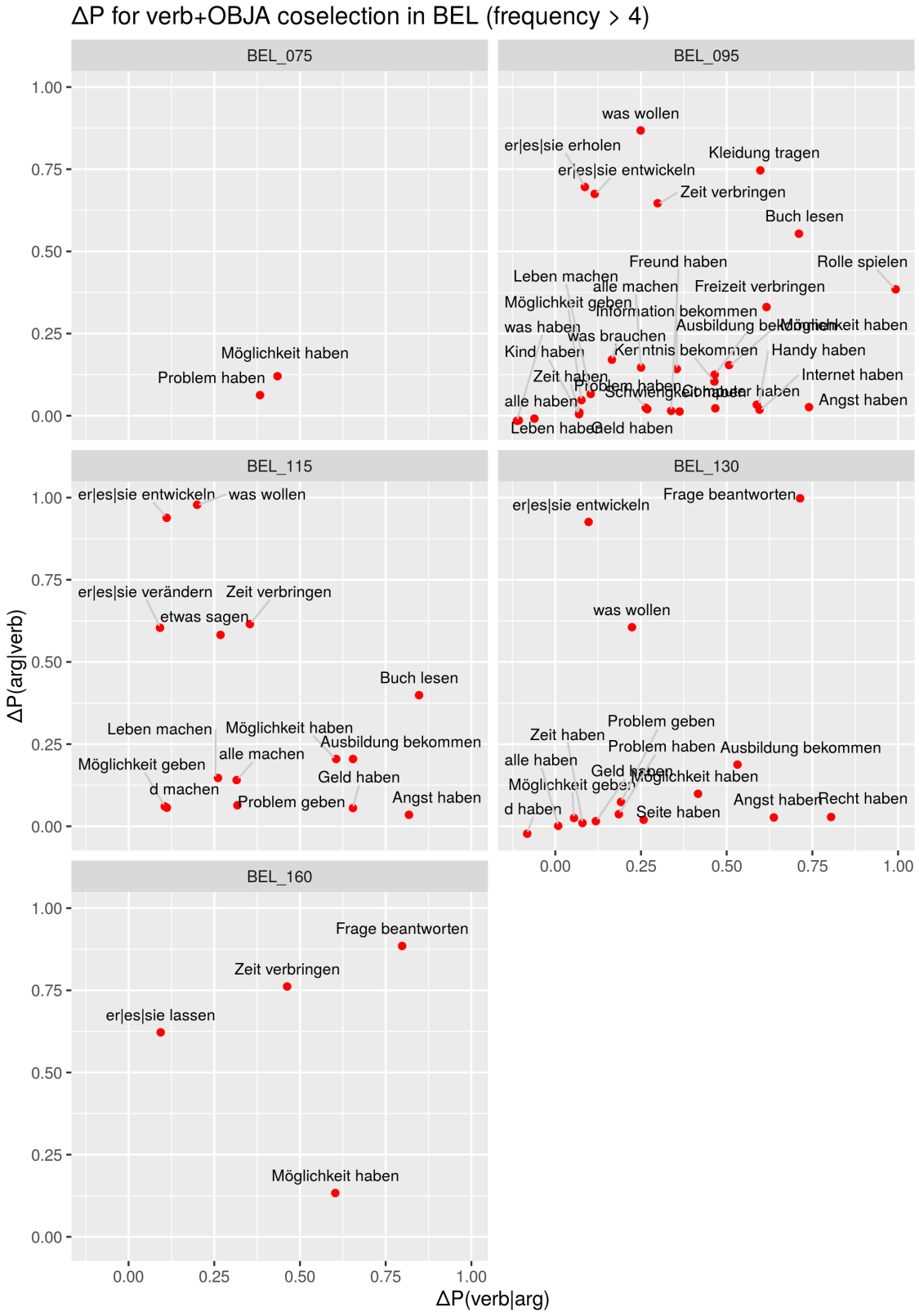


Figure 4.21.: ΔP for V+OBJA coselection in BEL, frequency ≥ 5

Figure 4.22.: ΔP for V+OBJA coselection in BEL, frequency ≥ 3

Figure 4.23.: ΔP for V+OBJA coselection in CH, frequency ≥ 5

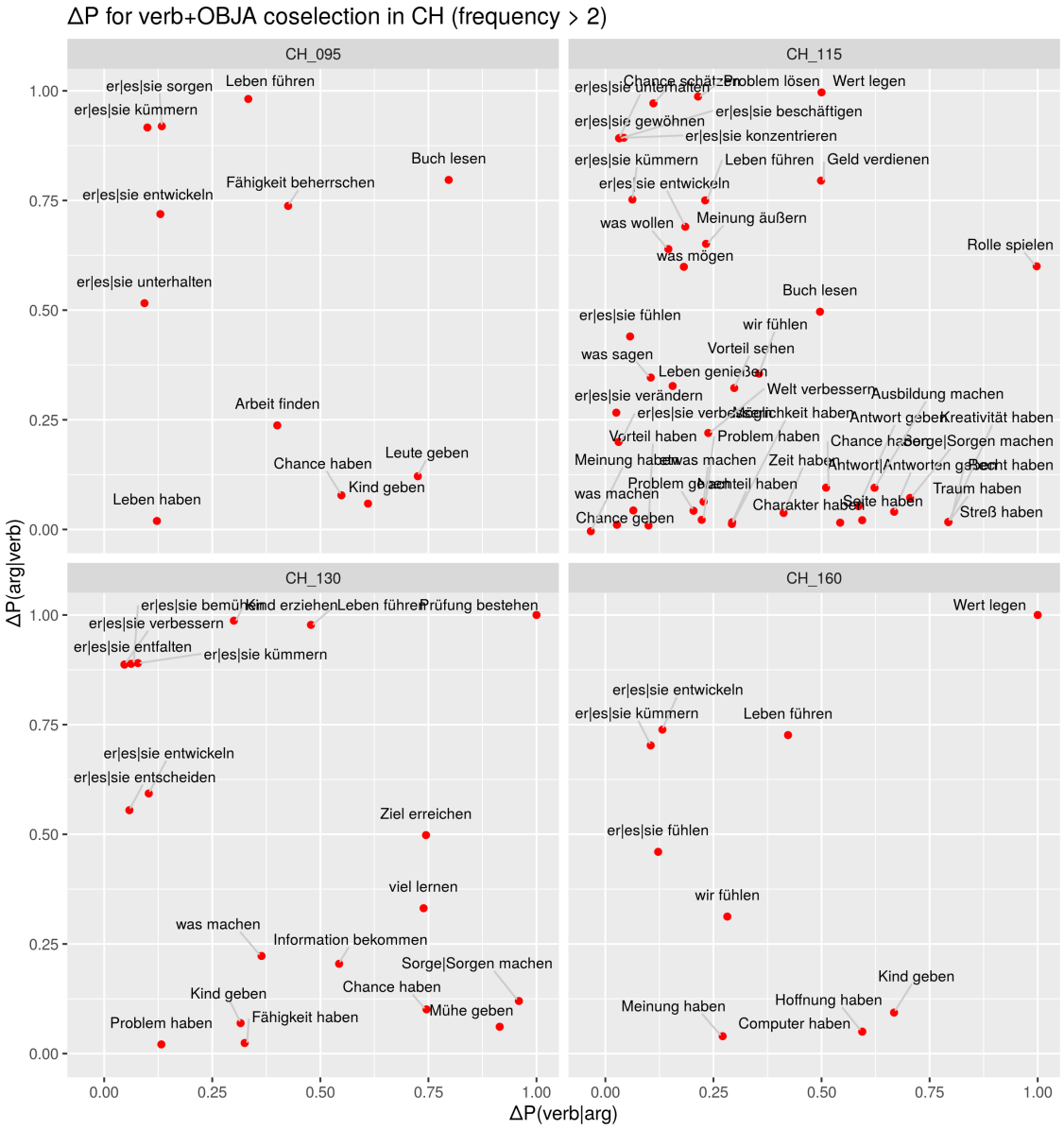


Figure 4.24.: ΔP for V+OBJA coselection in CH, frequency ≥ 3

4.3.2.2. ΔP(SUBJ)

For subjects, ΔP results fill more of the space of the measure. Weak repelling forces of *sein* ('to be') and some personal pronouns are visible in most plots. This is mainly due to the high frequency of *sein* in all subcorpora. Frequent coselections of V+ SUBJ are typically coselections of semantically light verbs such as *sein, haben, bekommen* ('to be', 'to have', 'to get') in both learners and native speakers. For L1 there is some variation in subject selection, such that more concrete or abstract nouns appear in subject slots than they do in learners. Those, however, are mostly prompt-related (*Jugendliche, Kind, Generation, Jugend*, 'adolescents', 'child', 'generation', 'youth'). The less frequent identical coselections do include some more interesting verbs, but subject fillers still heavily rely on either personal pronouns or prompt-related terms (see figs. 4.25 and 4.26).

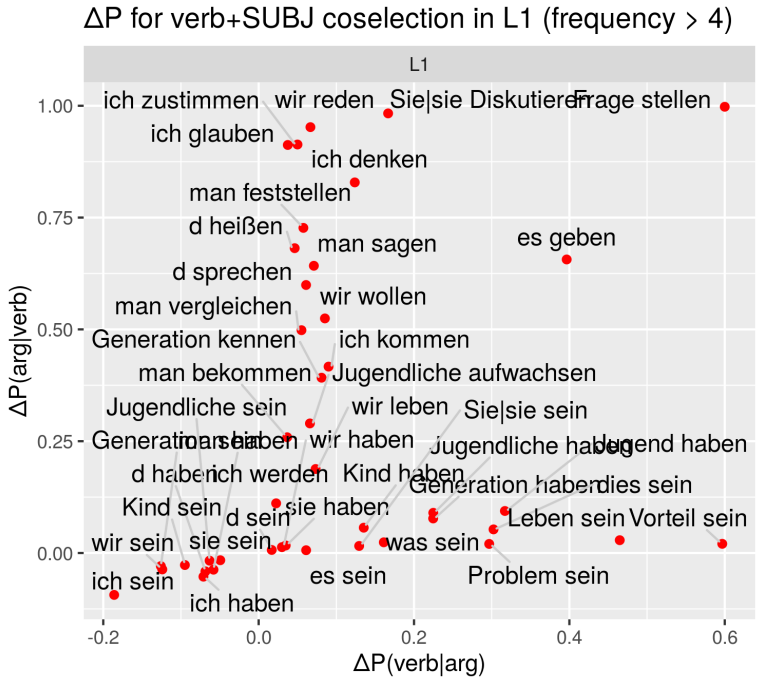


Figure 4.25.: ΔP for V+SUBJ coselection in L1

In learners, subjects are highly repetitive, much more so than in L1 (more on this in chapter 6.2) and rely almost exclusively on personal and indefinite pronouns and prompt-related nouns. In this section, results for learners are only shown for coselections of frequency 5 and higher for better legibility, results for coselections of lower frequency can be found in the repository (`10.5281/zenodo.3584091`). The highest number of identical coselection with a frequency of five or more exists in the BEL_095 and the CH_115 corpus, which are the largest corpora in terms of both documents and tokens included in the respective languages. It seems that for V+SUBJ, corpus size is a good predictor of number of identical coselection. This is not the case for all slots, as will be shown further below. It also seems that for V+SUBJ coselection, neither absolute or relative frequency nor ΔP hold particularly interesting insights. Viewing combinations of *haben, sein, bekommen* and personal and indefinite pronouns as coselectional constraints might be a bit of a stretch. It has been mentioned previously that pronouns are included in the analysis here, which in terms of statistical computation is not ideal (they skew values

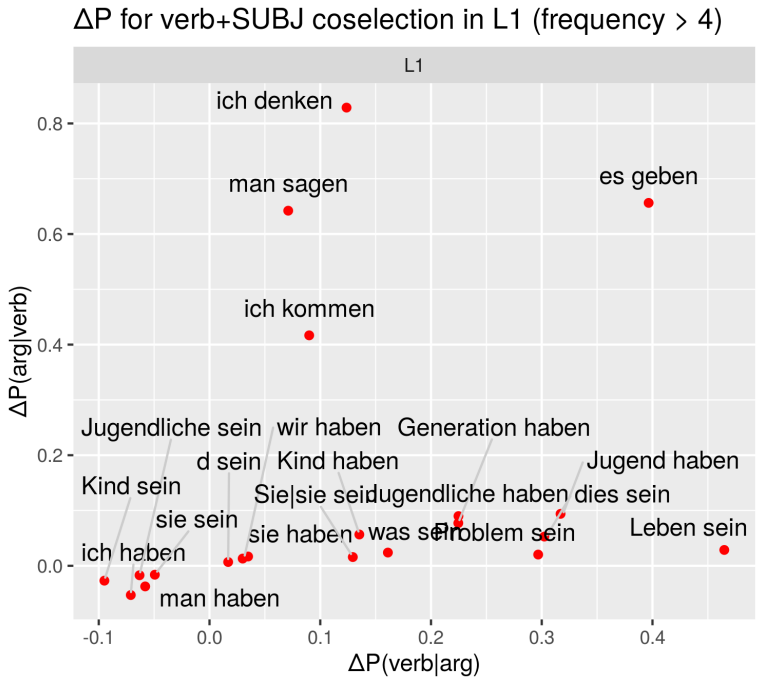


Figure 4.26.: ΔP for V+SUBJ coselection in L1

for lexemes of lower frequency, which are most lexemes) and with respect to semantically guided coselectional constraints as suggested by Plank (1984). An argument in favor of including pronouns is the discussion of coselectional constraints and preferences in more form-oriented and phraseological approaches. In those, pronouns could exhibit idiosyncratic or even systematic (distributional) coselectional preferences towards certain words and vice versa despite their semantic genericity.

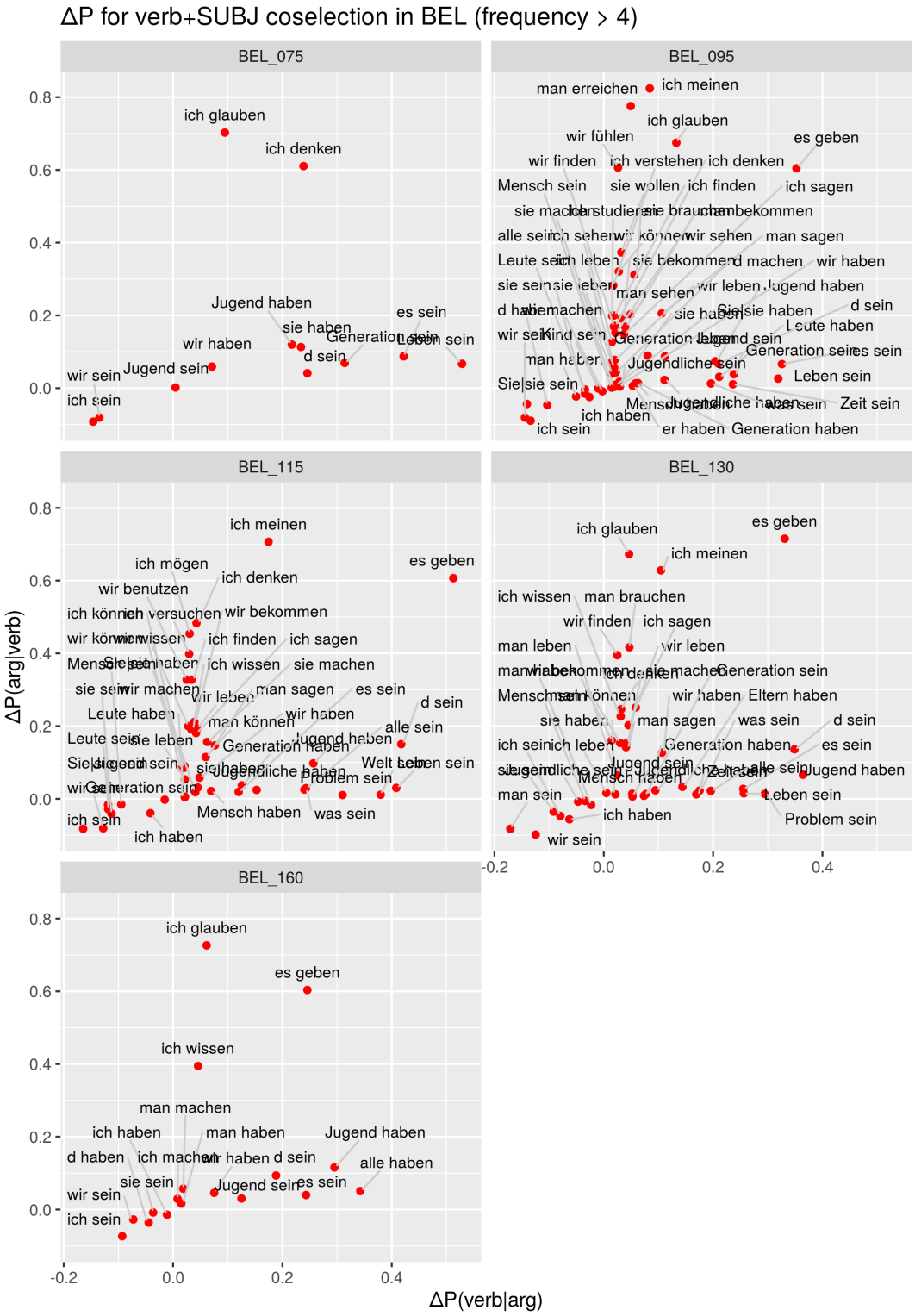


Figure 4.27.: ΔP for V+SUBJ coselection in BEL

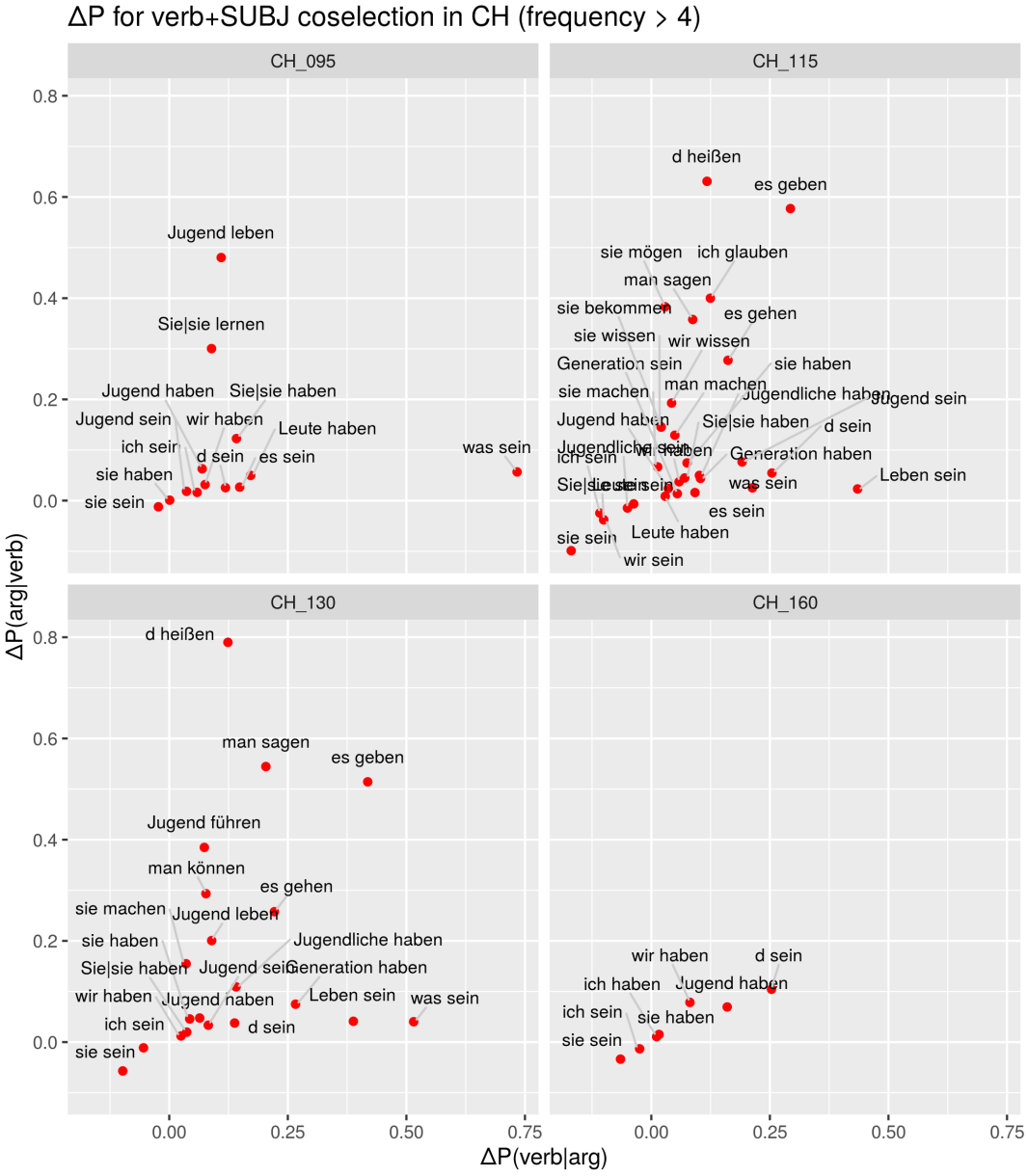


Figure 4.28.: ΔP for V+SUBJ coselection in CH

4.3.2.3. ΔP(PRED)

Regarding predicates, the ΔP analysis shows a striking qualitative difference between learners and native speakers:

- L1 predicates are more constructional (*es gut, leicht, schwer haben*, 'to have it good, bad, easy'), more discursive (*Die Frage, der Grund, das Problem, der Fall sein* ('to be the question, reason, problem, case') and more epistemic or existential (*möglich, aktuell, vorhanden sein* 'to be possible, current, to exist').
- In CH_095 and CH_115, predicates are much more descriptive (*fröhlich, glücklich, kreativ, intelligent, selbstständig sein* 'to be joyful, happy, creative, intelligent, self-reliant') or judgments (*wichtig, schlimm, schlecht sein* 'to be important, terrible, bad'). Some epistemic uses also appear towards the higher acquisition stages (*unvorstellbar sein, unmöglich sein, undenkbar sein* 'to be unimaginable, impossible, unthinkable'), but are outnumbered by the other uses.
- In BEL, there are many more predicates that appear identically in subcorpora, which is interesting given that in Belarusian and Russian, predicates are much closer to adjectives or modifiers in category because they are not connected with copula verbs and cannot always be told apart from postpositional adjectives. Their use of predicates seems descriptive, but in a different way from the CH learners, which is reflected in an extensive use of participles some of which can be read as deverbal predicative adjectives or state passives: *verbunden, verschmutzt, verboten, verändert, verwöhnt, vorhanden sein* ('to be connected, dirty, forbidden, changed, spoiled, to exist'), for a discussion on categorizing participles see section 3.2.2, and some less frequent adjectives such as *kompliziert, barmherzig, zielbewusst sein* ('to be complicated, merciful, purposeful'). At the same time they use more of the L1-like discourse-orienting predicates: *Die Frage, die Aufgabe, das Problem sein* 'to be the question, task, problem'); and also some predicates that point towards generalized statements: *die Zeit, das Leben, die Jugend, der Mensch sein* ('to be the time, life, youth, human').

In this way, the *types* of predicates in the languages and some of the differences between acquisition stages – particularly the decreasing number of identical coselections over time – is an interesting finding pointing towards stylistic and/or typological differences. However, the exact ΔP values seem somewhat random, which is an artifact of the low number of verbs that take predicates. They are basically only *sein* and *werden* ('to be', 'to become), and some verbs that require an adjectival argument or a resultative (*sich gut/schlecht fühlen* ('to feel well/unwell'), *es leicht/schwer haben* ('to have it good/bad'). This means that most of the values are more or less measured based on the frequency of *sein* and is also reflected in that in all plots, *gut sein* ('to be good') and *gut gehen* ('to be (feel) well') are in oppositional corners because *sein* appears with all kinds of other words, and *gut* appears mostly with *gehen*. Both are frequent, but due to a strong task- or prompt-effect, *gut sein* ends up in a corner where the two lexemes quantitatively repell each other.

Similarly to the OBJA-slot, results for ΔP(PRED) are *qualitatively* interesting, but quantitatively not interpretable.

4.3.2.4. ΔP(OBJP)

For OBJP, which is the slot underlying most lexicalization because many support verb constructions are based on OBJP and prepositions in OBJP are typically semantically

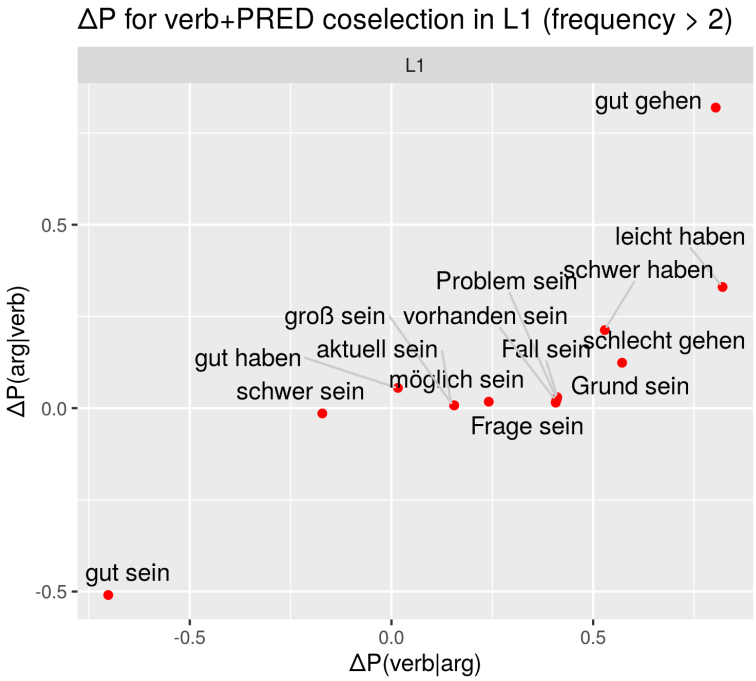


Figure 4.29.: ΔP for V+PRED coselection in L1

bleached, ΔP results are also quite interesting. First of all, it is obvious that there are even fewer identical coselections of OBJP complement slots.[22] Even if all coselections are included that occur only twice, there are only 11 in L1 (fig. 4.32):

- Some of those are are support verb constructions (*Funktionsverbgefüge*), in which verb semantics fade against noun semantics and some syntactic restrictions apply[23]: *Zur Verfügung stehen, zur Folge haben, an Bedeutung gewinnen* ('to be at the disposal (of)','to be consequential (to)', 'to gain importance')
- However, *an Freiheit gewinnen* ('to gain independence'), *auf die Jugend zählen* ('to count on youth', and *auf Geschwister aufpassen* ('to look after siblings') are less clear in terms of their phraseological weight, and *gegen ein Problem kämpfen* is arguably even somewhat unidiomatic.[24]

So it seems that while in OBJP, which should be the most coselectionally constrained category, because prepositions in OBJP are often semantically bleached and therefore inviting for support verb construction building, is in fact one where identical coselections appear more rarely than in OBJA or SUBJ. Of course they are also overall rarer, but only slightly rarer than predicates, yet still they show higher diversification. Perhaps this is because the OBJP slot, similarly to particle and prefix verbs, is less generic in meaning compared to other argument slots, and thus is used to convey specific information that is

[22] The complements plotted here are technically complements to the PP, not the verb.

[23] (?*Es wurde zur Folge gehabt*, literally 'It was had as a consequence' as the infelicitous(?) passivization of *zur Folge haben*, 'to entail')

[24] At least a search in the German reference corpus DeReKo Leibniz-Institut für Deutsche Sprache (2019) yields only one result for the exact *gegen ein Problem kämpfen* and zero for *gegen Probleme kämpfen*. *Gegen dieses Problem kämpfen* returns three hits, but is still negligible against *zur Verfügung stehen* (128 688 hits) and *zur Folge haben* (19780 hits).

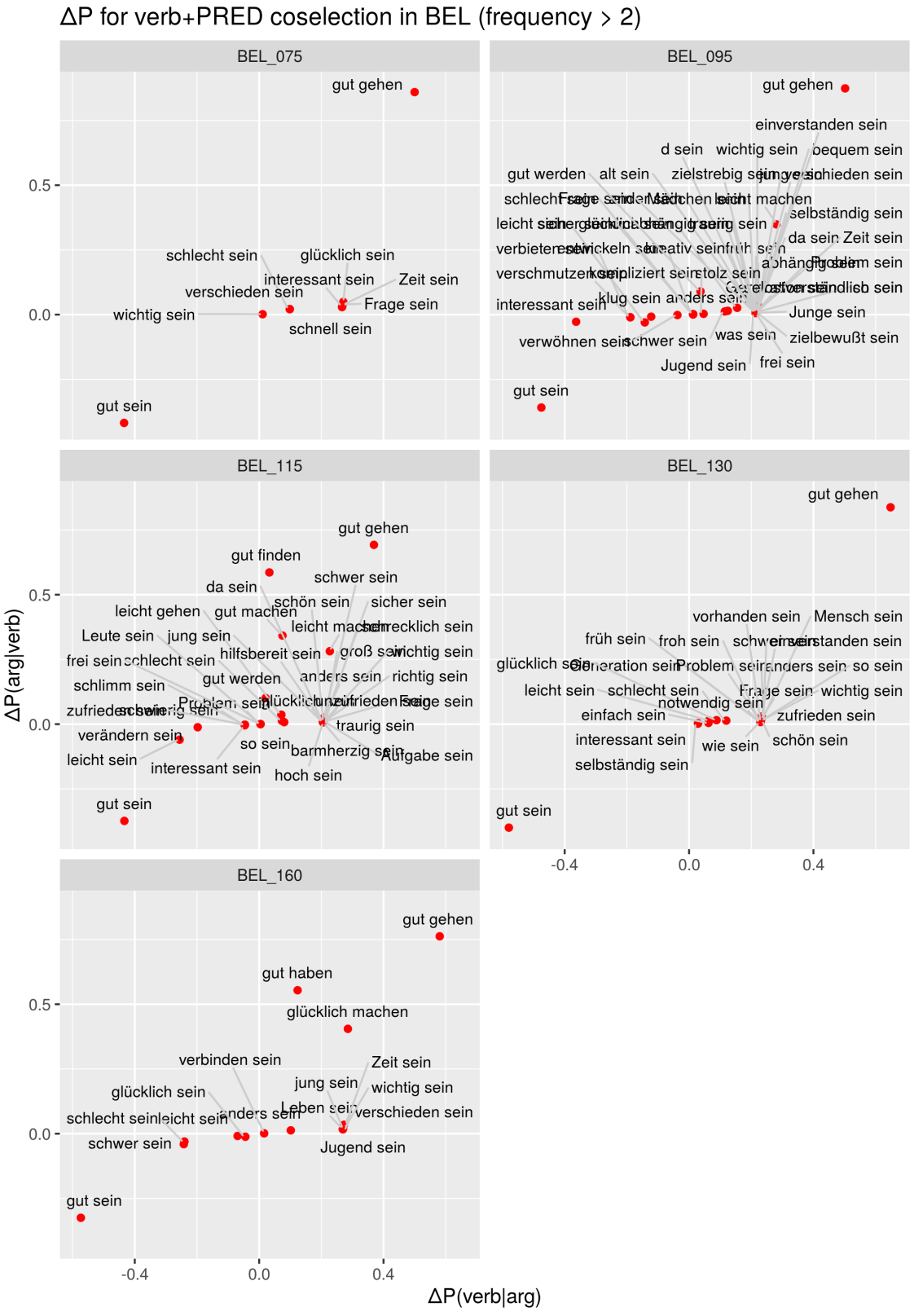


Figure 4.30.: ΔP for V+PRED coselection in BEL

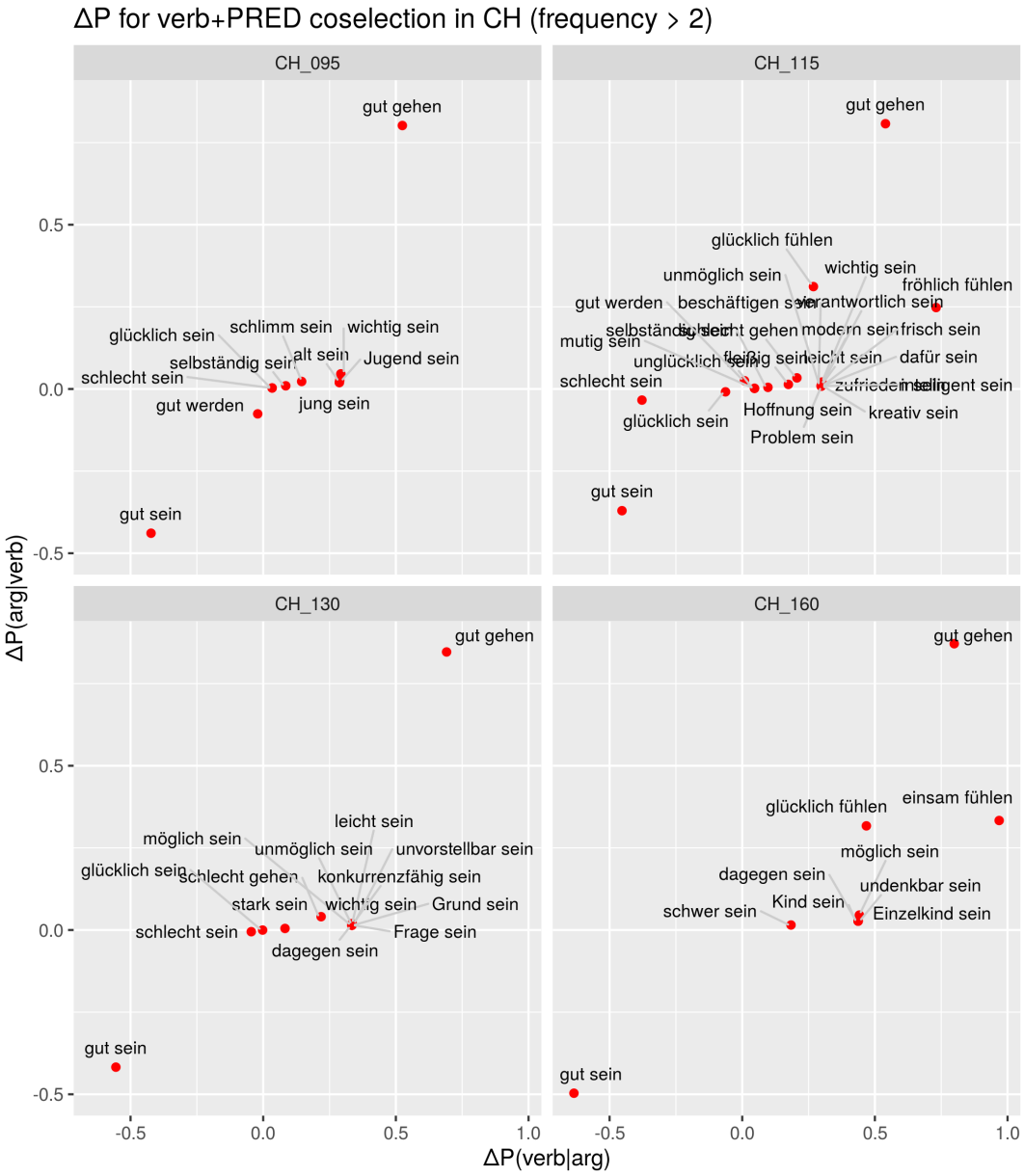


Figure 4.31.: ΔP for V+PRED coselection in CH

unlikely to re-occur as often. This would suggest that a specialization is problematic for the quantification of coselectional constraint, which is consistent with observations from the other slots.

This is further corroborated by the learners' OBJP coselection, where the hypothesis regarding a more similar writing at beginning stages and a higher diversification towards later stages can be confirmed. Both the BEL and the CH group use more identical coselections at earlier stages, and fewer at later stages. On a more qualitative note, the only somewhat frequently appearing coselection is *zur Verfügung stehen* and there is no clear trend in terms of abstractness or non-compositionality towards higher stages of acquisition:

- In CH-115 and CH-160, both *in die Schule gehen* ('to go to school', where the

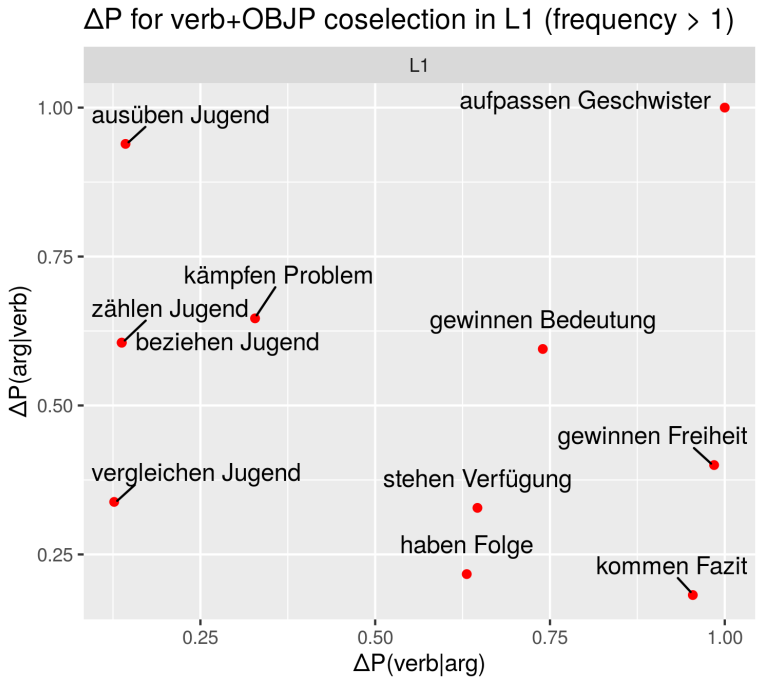


Figure 4.32.: ΔP for V+OBJP coselection in L1

directional argument was labeled as an OBJP, see sections 3.2 and 5.2 for details) and *unter Hunger, Schwierigkeiten leiden* ('to suffer from hunger, hardships') appear, and those mark two out of four repeating coselections in CH-160.

- In BEL-95, the number of identical coselections is the highest of all subcorpora, and most are highly concrete actions with directional objects. In the other BEL subcorpora, some more abstract coselections do appear (*in Betracht ziehen, im Stande sein*, 'to consider', 'to be capable of'), but identical coselections become rare overall. This can also be seen as evidence of a diversification of lexical material and topic which, again, makes it difficult to track the development of coselectional preferences or constraints in an exemplar-based fashion.

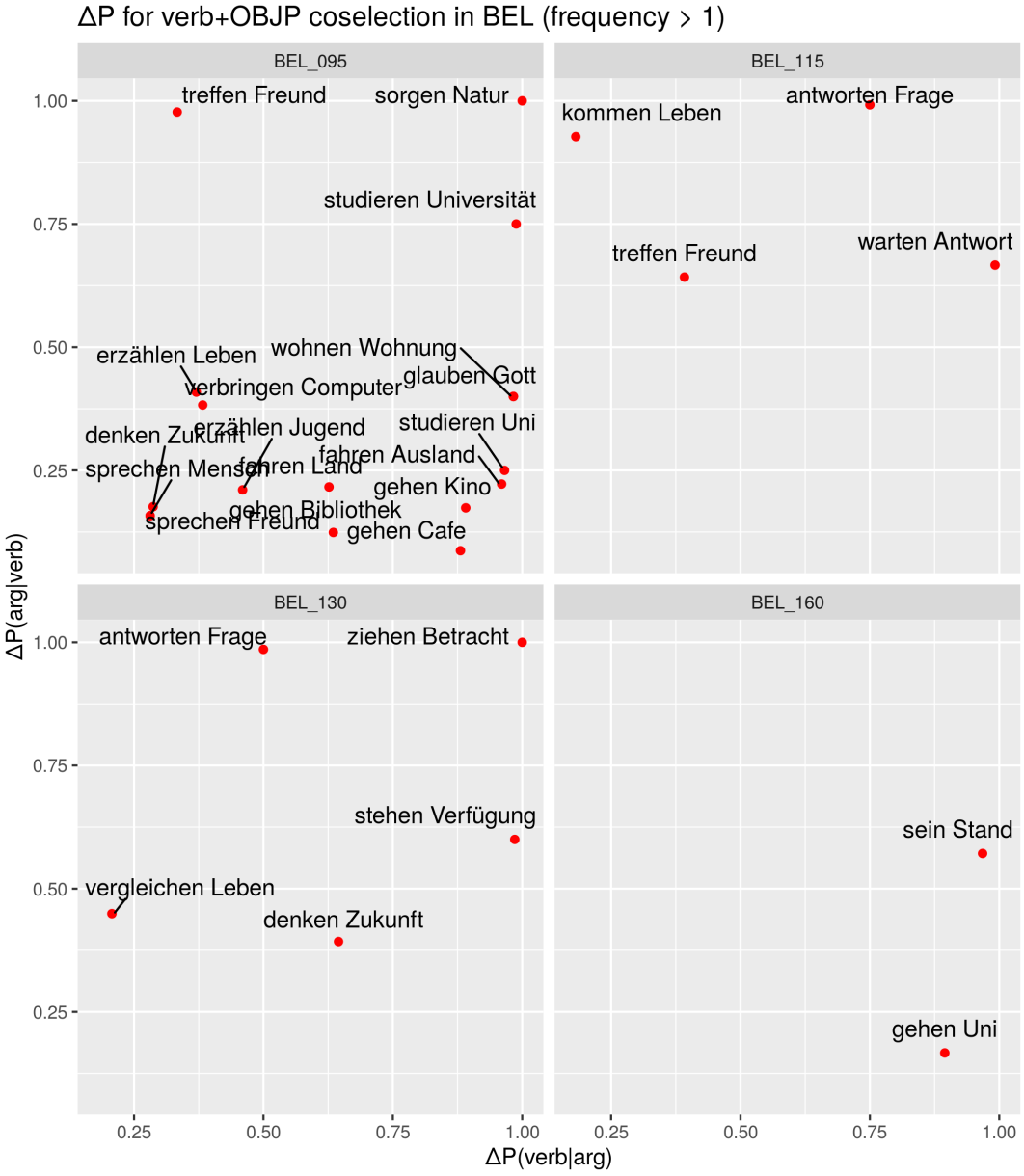


Figure 4.33.: ΔP for OBJP coselection in BEL

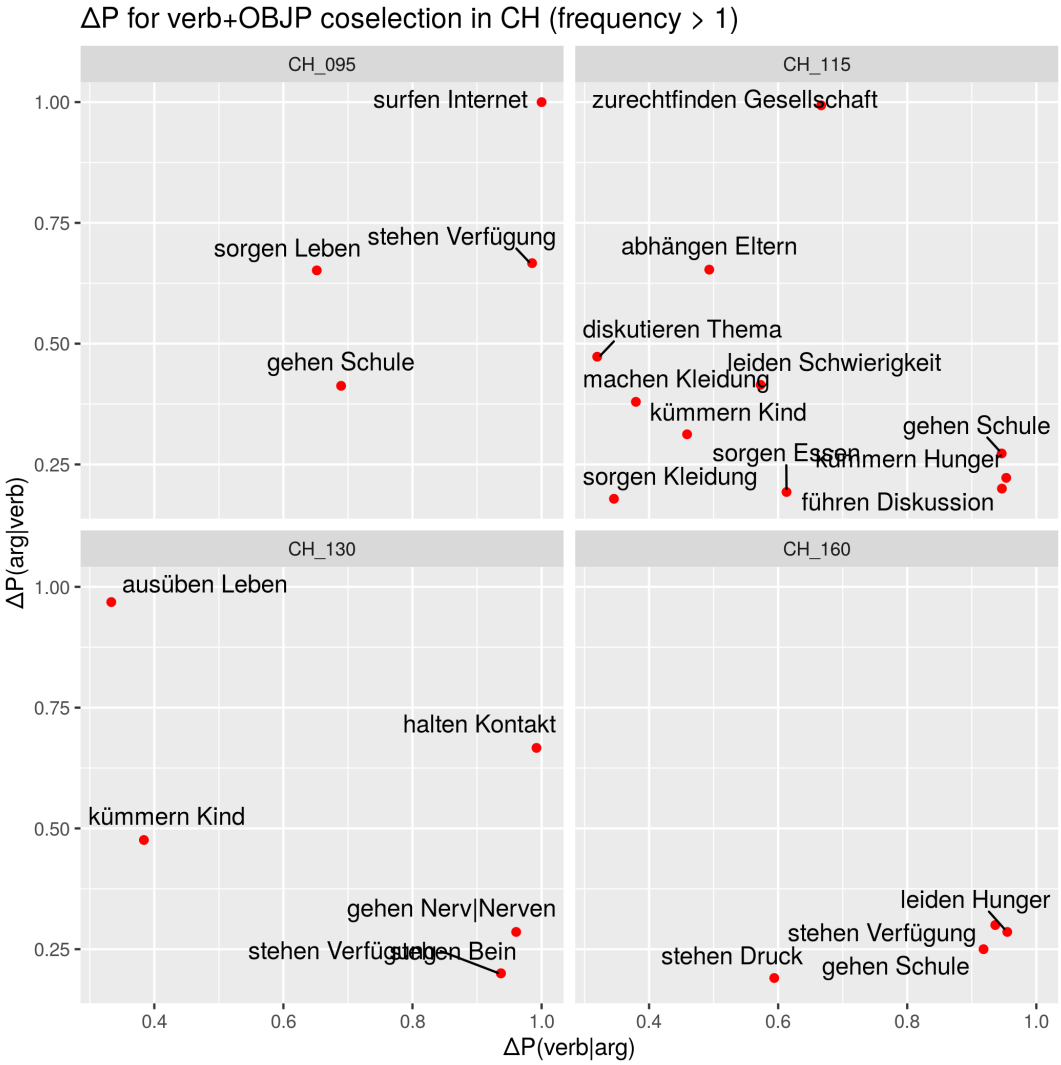


Figure 4.34.: ΔP for OBJP coselection in BEL, BEL_075 does not have identical coselections for OBJP

To summarize the results from this subsection:

- For the OBJA slot, ΔP values work as a categorizer into more verb- and more noun-driven coselection (reflexive verbs vs. more lexicalized and noun-driven combinations like *Entscheidung treffen, Recht haben*, 'to take a decision', 'to be right'). ΔP plots from later onDaF stages in learners structurally resemble L1. The scalar nature of the measure does not materialze in data of this size due to floor effects with respect to the expected frequency of occurrence of coselections.
- For subjects, ΔP does not seem to yield particularly interesting results.
- For predicates, differences between language groups and acquisition stages are interesting, but the exact values are not very telling due to their high dependence on the high-frequency verb *sein* ('to be').
- Perhaps the most interesting results exist in OBJP. But even there, it is the distribution across categories that is most interpretable, while the number of identical coselections is so low that the validity of the values is at least questionable.

So it appears that ΔP with its two-dimensionality offers a good view on distributional differences and even works as a morphosyntactic and semantic categorizer, but it does not provide a framework for an estimation of the development of coselectional constraint as a structural category. Although some epistemological concerns regarding the measure itself have been raised, it seems that in this case, the issue is actually a deeper one. It appears that an item-based analysis, even if it relies on the distribution of the other items, is simply not well suited to capture the effects of coselectional constraint at least in corpora of this size.

4.4. Summary

Descriptive statistics of Kobalt show meaningful and systematic differences between L1 and L2 and between two L2 groups in lexical, morphosyntactic, and syntactic distributions. Developmental trajectories for lexicosyntactic categories in L2, clear L2 vs. L1 effects and within-group effects for the two learner groups (BEL and CH) were found, but also considerable variance in L1 and L2.

It appears in several statistics that CH is more similar to BEL lexically and more similar to L1 syntactically, and that CH and L1 are generally more similar to each other than BEL and L1. In some statistics, CH is even closer to L1 than to BEL, suggesting that aside from general L2 effects, typological and teaching effects may also play a role in distributional aspects of lexicosyntax. Some aspects in the development of the verb-argument structures and verb categories also point towards an increase in coselectional constraint, like the growing number of prepositional objects or the construction verb category *cx*.

At the same time, absolute frequencies for individual lexemes or combinations of verb argument structures quickly disperse into very low numbers. This means that identical items cannot be traced across corpora in nearly any cases.

In line with the prevalent methodology of quantiative phraseological research, a lexical association measure was computed for the OBJA, SUBJ, OBJP, and PRED slots in all subcorpora in order to gain an understanding of coselectional constraint in each corpus; and in search of promising candidates for further inspection. However, while yielding

interesting qualitative results, the ΔP analysis did not provide an obvious way forward for a quantification of the structural extent of coselectional constraint.

Given the small size of the corpus and the low number of exact re-occurrences of co-selected items it is in fact surprising that some interesting distributional observations can still be made, such as a higher overlap in earlier BEL corpora (despite shorter text length), differences in categories or types of strongly associated co-occurrences between learners and L1, such as in the OBJP and PRED slots, or changes in the SUBJ and OBJA slots in the most advanced groups.

The main compromising factor for a statistically based quantification appears to be the predicted process of progressive lexicosyntactic diversification, rearrangement, and specialization. Evidence for such a process has been observed in the development of TTRs, the pairwise intersection of lexemes between all texts divided by language and onDaF groups, the list of most frequent verbs, the diversification of verb annotations, and in ΔP measures. While in line with the predictions, it also renders the assessment of how individual items develop impossible, because they either disappear or only start appearing in corpora of late onDaF ranges. Persisting items are mostly functional or prompt-related (which means they may not be coselectionally constrained to the same extent outside of the prompted context). While functional items may also underly coselectional restrictions, they can be functionally diverse and some do not take arguments in all their word senses (like *sein*, 'to be', or *haben*, 'to have', in auxiliary function vs. use in support verb construction or modal infinitive).

Despite the prediction of a process of diversification and specialization, its problematic effect on the quantitative analysis was underestimated, especially of a smaller corpus. Most of the recurrent vocabulary in Kobalt is made up from either topic-related, functional, or rather general lexemes, and the rest are hapaxes that are not usable in a quantitative analysis. An operationalization of the structural development of coselectional constraint as discussed in section 3.1 can therefore not be deduced from the statistics as they are discussed. It seems doubtful that an approach focussing on concrete items is capable of capturing the expected structural changes.

Another problematic aspect of a statistical operationalization, aside from the epistemological concerns raised in section 4.3.1, is that it would require a triangulation of a number of measures rather than providing a single measure of comparison.[25] This is not a well-researched area in linguistic methodology and would require more tentative modeling, and likely still result in a significant degree of uncertainty. In conclusion, a statistical and item-based approach does not seem to capture the effects well, in spite of some hints at their existence, and thus does not provide viable framework for the study of the research question in this data. If the study is to remain quantitative, a different approach should be developed, one that requires additional information. This, as shall be seen in the next chapter, need not necessarily stem from additional data.

[25] This is also what Gries (2019, 396) suggests as a solution for lexical association measures in general, labeling it "the tupleization of corpus linguistics, namely (i) the collection of multiple values per event type, where event type can refer to an individual element or, more the focus here, the co-occurrence of elements and (ii) the use of as many of those values as possible in the analysis/interpretation part". This, however, would not work as a measure for comparison, unless the elements of the tuple were contrasted against one another individually, which is done by Gries in three-dimensional plots. A visual analysis is limited to three dimensions and often unreliable. Still, his examples seem somewhat more revealing than a simple ranking by lexical association measures on a single dimension or two, like in ΔP, but not as a solution that captures the full complexity of coselectional preferences for the total distribution. It would also not resolve the problem of disappearing items; nor would it provide an operationalization of the concept of coselectional constraint *as such*.

5. A graph-based model of verb-argument coselection in Kobalt

The description and analysis of the data so far has relied on counting individual items and individual item combinations in relation to the frequency of other items in the corpus. What cannot be accessed and used in frequentist or probability-based approaches relying on factor combinations is the relationship between items, such as all verbs and all arguments, and their connectivity. This can, however, be modeled in a graph, providing a layer of information that is much richer than just the frequency of co-occurrence and will be discussed in this chapter.

I will first introduce graphs as a data structure and describe their current use in linguistics, computational linguistics, and digital humanities (DH), showing also that graphs are informationally denser and therefore beneficial for extracting quantifiable information from small and mid-sized corpora. I will then present a graph-based model of verb-argument coselection in Kobalt at different levels of specificity and finally introduce a quantification of graph structure, namely Louvain modularity (Blondel et al., 2008), that will be used as a measure for coselectional constraint in Kobalt. Results and a methodological validation will remain for the next chapter.

5.1. Graphs as a data and knowledge structure

Graphs are a type of data structure that centers on the relations between entities in the form nodes (vertices) and edges between them. Formally, graphs are two-tuples of unordered sets, $G = (V, E)$, where

$$V = \{v_1, ..., v_n\} \text{ and } E(G) = \{e_1 = (v_{source}, v_{target}), ..., e_n = (v_{source}, v_{target})\}$$

in a directed graph and

$$V = \{v_1, ..., v_n\} \text{ and } E(G) = \{e_1 = \{v_{source}, v_{target}\}, ..., e_n = \{v_{source}, v_{target}\}\}$$

in an undirected graph. The two differ only in that v_{source} and v_{target} may be swapped in the undirected graph, which is to say they are contained in a set (marked with braces: {}), not a tuple (marked with round brackets). Technically, they thus are no longer distinguishable as *source* and *target* nodes. In a property graph, both the nodes and the edges can have any number of properties, and edges can signify any kind of relation between two nodes. They can be weighted or unweighted, i.e. modeled to possess greater importance or frequency, and directed or undirected. Directed edges signify a one-sided relationship, such as inheritance, while undirected edges signify a mutual relationship, such as genetic relatedness.

Graph theory is a branch of discrete mathematics that focusses on the classification and analysis of graphs based on their abstract properties, like isomorphism (structural identity), the (im-)possibility to divide a graph into several subgraphs based on their properties, the extraction of the largest interconnected subgraph (clique, largest component),

the (im-)possibility to color or walk through the graph in specified ways (Scheinerman and Ullman, 2011), distance between nodes and the size of the graph as defined by different measures (such as the shortest path), and algorithms based on that. One of the most-cited introductions is Golumbic (2004). Mihalcea and Radev (2011) provide an NLP-oriented introduction of graph theory listing some of the algorithms and applications based on graphs that are used in computational linguistics and NLP today.

The discipline of *network theory* or *network analysis* looks at the functional aspects of subgraphs and nodes, such as the relative importance of certain nodes and edges in the graph, and their development over time.[1] In the words of Borgatti and Halgin (2011, 1168),

> "Network theory refers to the mechanisms and processes that interact with network structures to yield certain outcomes for individuals and groups. (...) [N]etwork theory is about the consequences of network variables, such as having many ties or being centrally located".

They add the term *theory of networks*, which

> "refers to the processes that determine why networks have the structures they do (...). This includes models of who forms what kind of tie with whom, who becomes central, and what characteristics (e.g., centralization or small-worldness) the network as a whole will have".

In this sense, theory of networks is not so much a set of theories of networks, but a set of predictions of the world as seen through the lense of specific subjects modeled in graphs. Network theory or network analysis then is situated at a lower level of abstraction regarding the inner dynamics of a specified graph, focussing on the concrete dynamic processes of the graph as a representation of a specified system (for example a graph model of the lexicon). Theory of networks relates to network-specific, but abstract dynamic processes such as growth functions and their dependence on properties of networks *as such*. Graph theory as a topological discipline does not take a functional view of subject-specific graphs, but classifies graphs by characteristics of their nodes, edges, paths, isomorphisms, ability to be partitioned or transformed, etc. However, in practice, there exists some terminological confusion, with an addition of another two terms, *theory of graphs* (Scheinerman and Ullman, 2011; Mesbahi, 2002) and *network science*. All five terms – graph theory, theory of networks, network theory, theory of graphs, and network science – are used as partially overlapping or even interchangeably.

Models based on (abstract) graphs and the analysis of (concrete) networks have played a massive role in all STEM disciplines,[2] but most obviously in computer science, where graphs are used as a data structure for storage and retrieval (databases and search algorithms and engines), as data models (process flow design), as data formats (trees, e.g.

[1] Examples include ranking metrics for efficient web search such as HITS (hubs and authorities, Kleinberg (1999)) or PageRank (Page et al., 1999), versions of in-/out-degree centrality as often used in citation studies in political science and jurisdiction studies (Lupu and Voeten, 2012; Ighreiz et al., in prep.; Coupette, 2019), measures of modularity or connectivity such as Louvain modularity, which will be used in this study (Blondel et al., 2008), and measures of assortativity (likelihood of nodes to connect to nodes of similar degree), among others.

[2] Examples include models of properties of electrical circuits (Harary, 1959), transit networks (Derrible and Kennedy, 2011), evolutionary dynamics (Lieberman et al., 2005), solutions to master equations in physics (Schnakenberg, 1976), fMRI and MEG neuroimaging (Mandke et al., 2018), geostatistical modeling (Tahmasebi and Sahimi, 2016), and topological aspects of organic and inorganic chemistry (Trinajstic, 2018).

XML), and algorithmically for many tasks, such as memory allocation (Callahan and Koblenz, 1991), task management and distribution (Chandy and Misra, 1982; Gonzalez et al., 2012), network design, routing, and security (Krumke and Noltemeier, 2005; Lazos et al., 2005).

In computational linguistics and natural language processing (NLP), graphs are used for a wide range of applications, including, but not limited to text mining, e.g. word sense disambiguation for classification (Veronis and Ide, 1990; Rousseau et al., 2015) or word centrality for text summarization (Erkan and Radev, 2004), text generation (Krahmer et al., 2003); and structural analysis, for example in syntactic and morphological parsers (Woods, 1970; Wittenburg, 1986; Nivre, 2004; Nivre et al., 2006; Seeker and Çetinoğlu, 2015), machine translation (Ueffing et al., 2002; Alexandrescu and Kirchhoff, 2009; Bastings et al., 2017), or random forest training in machine learning and other applications (Palomino-Garibay et al., 2015; Kobyliński and Przepiórkowski, 2008; Xu and Jelinek, 2004).

Since the focus of computational linguistics currently lies on the development of applications for information extraction and language-based human-computer interfaces, graphs are in most cases used as databases or as intermediate steps in analysis pipelines rather than as research objects in their own right.

More recently, the digital humanities (DH) have adopted graphs as a more common way of representing knowledge in so-called knowledge graphs (Haslhofer et al., 2018) or for network analysis similar to applications from more sociological fields, for example with networks of co-appearance in dramatic works in literary studies (Kuczera, 2017) or acquaintance and citation networks in a variety of historical studies.[3] One of the most influential recent ideas in corpus linguistics and related disciplines across the digital humanities is the concept of *text as graph*, where nodes represent elements on all levels of granularity, beginning at character level (or phonetic units in the case of spoken language) and then moving upward towards syllables, tokens/words, sentences, paragraphs, and all the way up to chapters, books, collected works of one or many authors and sublanguages from a period or a region and so on. Each level of granularity includes the nodes from the lower levels, which are linked with ordering edges labeled for example as 'preceeds' or 'follows' on at least one of the levels.

A representation of text as a graph can be favorable over text as XML (tree)[4] or string, because the nodes in a graph, unless otherwise specified, are not ordered, and therefore allow for a break-up or overlap of hierarchies on various levels – there is no need to *uniquely* specify "containment, dominance (hiearchy), datatyping, and order" (Haentjens Dekker and Birnbaum, 2017, section Markup). Rather, text modeled as a graph allows for a simulatenous categorization on different levels through unordered edges: Edges can connect tokens directly to texts without first going through paragraphs, sections, chapters, and so on, or can build one path going through those layers and one that avoids one or more of them. This also means that text as graph provides space for directly connecting discontinuous elements, such as interrupted speech or the simple case of German particle verbs,[5] without breaking up or doubling the tokenization layer.

[3] See Ahnert and Ahnert (2015); During (2016); Wilcke et al. (2018) and a bibliography on `http://historicalnetworkresearch.org/bibliography` for a number of examples.

[4] A tree is an acyclic graph where any two nodes may only be connected by one path. This means there can be no edges between nodes of the same layer in a hierarchy, and only one edge connecting a node to the next higher level of hierarchy. While trees are graphs, *text as graph* refers to more flexible graphs than trees.

[5] Complex German verbs that incorporate a particle such as a preposition (similar to English phrasal

On the more technical side, text as graph is a way to model, represent, store and search language data and is natively represented in graph databases such as neo4j (`www.neo4j.org`) or graphANNIS (Krause, 2019) where it can be connected with non-text entities, effectively providing an interface for searching across text and non-text, not only for filtering, but for specified search and detection of relationships between items that are not or do not appear to be connected on the surface, or are not labeled accordingly (concept-based search, see Efer (2017, chapter 3.7) for a DH-specific introduction). Linguistically specialized search algorithms based on prefiltered subgraphs in graphANNIS (Krause, 2019) also offer a strong speed advantage over more traditional binary search and filtering approaches (which are, despite their reliance on trees, not typically termed graph searches) in mid-sized and large corpora.

In conclusion, graphs are a powerful instrument in the exact and quantitative modeling of relational problems; and with a growing set of algorithmic implementations, metrics, and mathematical proofs they allow for wide-ranged and fruitful application in all computationally oriented fields. Since they are only beginning to find their way into core-linguistic research, the next section will provide a short overview of the specific properties of graphs that make them interesting for the modeling of language. It aims, on one hand, to motivate the use of a graph-based model and explain how it can avoid the problems of the statistical approach discussed in the previous chapter; and on the other, to contextualize the model against both the existing research landscape. Some more specific suggestions to this extent will be made in the discussion in section 7.2.

5.1.1. Graphs and linguistic theory

In linguistics, graphs have been used most widely and prominently in the form of syntax or constituency trees, which are a special type of graphs in that they are hierarchical and acyclic; and as lexical or semantic networks. In the case of syntax, edges usually signify dependency or constituency within a single sentence, phrase or subphrase. In the case of lexical networks, they usually signify co-occurrence within a certain text, word range, text type, or register. Earlier in the field, more attention was directed at graph-theoretic work and attempts to model linguistic problems in Markov Chains in particular, which are graphs in which nodes represent states and edges represent transitional probabilities, i.e. probabilities to reach one state from another one (Brainerd and Chang, 1982; Jelinek et al., 1975; Goodman, 1961). However, this line of work already started out closer to the computational and NLP subfields and quickly moved even further away from core linguistics. In effect, graphs and graph theory are rarely used for linguistic modeling and analyis. More commonly, they visualize existing analyses, i.e. represent the final result of research, ratner than providing ground for exploration or research questions specifically aimed at the relations found in a given dataset.

As far as pure visualization goes, the same co-occurrences can be represented as a combination of factors as in tab. 5.1 – which is the same representation as it was used for lexical co-occurrence statistics in chapter 4 – or as a network, sometimes referred to as a lexical co-occurrence network (Edmonds, 1997; Chen et al., 2018; Mollet et al., 2012), as in fig. 5.1 (example from Kobalt). Co-occurrence networks can rely on positional or syntactic co-occurrence, in this case edges represent accusative object dependency (OBJA).[6]

verbs). In their finite occurrence, the verbs split into two parts, the base verb and the particle: *Er schreibt sich das Rezept auf* ('He is writing down the recipe'); but in the infinitive, they occur as a single word: *Er hat sich das Rezept aufgeschrieben* ('He has written down the recipe').

[6]The frequency of co-occurrence is not modeled in this example but is considered in the model used in

verb	noun	frequency
behalten ('keep')	*Vorteil* ('advantage, upside')	1
bieten ('offer')	*Vorteil* ('advantage, upside')	1
bringen ('bring, provide')	*Vorteil* ('advantage, upside')	2
genießen ('enjoy')	*Vorteil* ('advantage, upside')	1
sehen ('see')	*Vorteil* ('advantage, upside')	4
suchen ('seek')	*Vorteil* ('advantage, upside')	1
suchen ('seek')	*Wahrheit* ('truth')	1
suchen ('seek')	*Freund* ('friend')	2

Table 5.1.: Some of the verbs that take *Vorteil* as an accusative object in Kobalt and their coselections

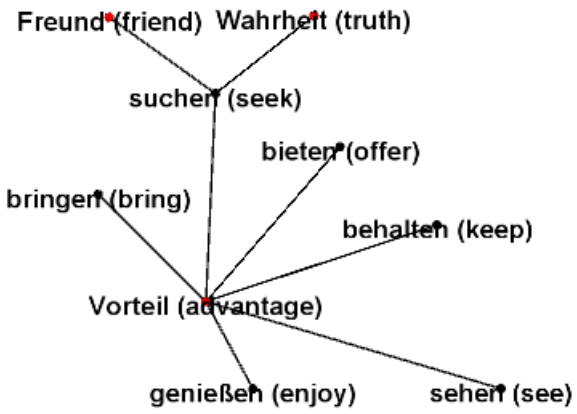


Figure 5.1.: Verbs and OBJA selections from tab. 5.1 as a lexical co-occurrence network. Graph visualized with Gephi (Bastian et al., 2009)

In terms of quick access of information, both present advantages and disadvantages: Relational or structural information can be accessed more quickly from the graph, while precise information, such as the frequency of co-occurrence, is more easily accessible from the table. In graphs which model frequency as edge weight, it can be hard to tell apart thicker edges precisely from thinner ones depending on the dimensions of the visualization. Often, a smaller graph tends to be more legible and provide quick access to relational information, while larger graphs tend to be harder to read than tables, because individual nodes are not as easily found unless they are very central graph elements.

But rather than being merely or primarily of representational merit, graphs are in fact their own information or knowledge structure and contain more explicit information than a list of factor combinations. Graphs *explicitly* model the assumption of an existing connection (or disconnection) between all nodes, as for example all different dependents of the same verb. Conceptually, this has already been suggested by CxG in construction or inheritance networks (Lasch and Ziem, 2014; Zeldes, 2013a; Michaelis, 2012) and the *Semantic Coherence Principle* (Goldberg, 2006). Graph-based models also account for this assumption in the quantitative analysis.[7] This adds another dimension to a graph when

the study as weighted edges, see section 5.2.

[7] A quantification of the slots of individual verbs or nouns has been provided in productivity studies such

compared to a table, not just on paper, but with respect to the amount of information that is condensed in the model. For each item (a factor in the first case, a node in the second), additional information is encoded in the graph, namely the existence or non-existence of a path between any two nodes. Efficient analysis of large graphs therefore still requires a relatively large amount of computational resources even with the computational power currently available. That explains in part why graph-based engines and analyses are only beginning to blossom outside of strongly computationally oriented subjects.

Moreover, graphs do not only make explicit the relationships between what are rows in a matrix of factor combinations in statistical analysis, but they also effectively abstract away from concrete elements and their features and allow for comparison of diverse element types *exclusively* through their relationships with all other elements regardless of their type or ontological status. A graph can be populated by all kinds of entities, elements or other objects, or simply vertices as understood by mathematical definition, points in a continuum specified only through their position in relation to the axes of the coordinate system.

In fact, there has been a discussion in philosophy/metaphysics that goes so far as to suggest that the world (meaning 'all of existence') cannot only be *represented* in a graph, but *is* in fact a graph, which is to say it can be defined only in terms of the relationships between other sets of relationships, such as the relationships between subatomic particles, which are understood as forces – i.e. relationships – to form atoms; which then form molecules through their relationships; the relationships of those molecules, which are then graphs themselves, to form larger 'entities' consisting of the smaller subgraphs and relating to one another in specific and specifiable ways and so forth (Dipert, 1997). In Dipert's understanding, the nodes of such a graph are not capable of 'possessing' features themselves. Instead, features are in fact subgraphs, i.e. relationships between yet smaller subgraphs. So both 'entities' and 'features' *are* rather than 'have', 'acquire', or 'lose' groups of properties. Features and entities in this model do not differ ontologically, they are merely smaller or larger subgraphs. It is the relation between those vertices that creates all that is, illustrating that graph-based modeling and reasoning – if understood as an ontological representation of a structure – in fact raises questions far beyond pragmatic issues of legibility, elegance of visual representation, or computational and representational benefits through a certain type of data structure.

This discussion may seem far outside of the present-day discourse in linguistics, but actually, fairly similar ideas have already been formulated in Halliday's continuum of lexis and grammar (Halliday, 1992, 63):

> "The point is that grammar and vocabulary are not two different things; they are the same thing seen by different observers. There is only one phenomenon here, not two. But it is spread along a continuum. At one end are small, closed, often binary systems, of very general application, intersecting with each other but each having, in principle, its own distinct realization. (...) At the other end are much more specific, looser, more shifting sets of features, realized not discretely but in bundles called "words" (..); the system networks formed by these features are local and transitory rather than being global and persistent"

and in Goldberg's "constructions all the way down" (Goldberg, 2006, 13).

as Zeldes (2012, 2013a), but a unified quantification of the whole network has not been suggested yet to my knowledge.

The idea that grammar and lexicon cannot be separated has been widely accepted in usage-based linguistics, but few models exist that would attempt to model their coexistence within the same space formally. What CxG most often implies is instead a gradual in- or decrease of productivity/fixedness and willingness to accept other elements into slots, with the most abstract constructions being those that are least specified lexically, and the most concrete ones those that do not allow for any changes. But this alone does not specify which items are to be found at which coordinates of the continuum and how and where they interact. In fact, 'continuum' is a somewhat misleading term altogether because usually linguists refer to several categories that gradually differ from each other in fixedness and that can be separated more or less clearly – this is impossible in an actual continuum such as the continuum of real numbers, where any two numbers cannot be separated clearly, no matter how small the difference is in terms of decimal places. In all its current implementations, construction grammar works with the idea of constructions and construction slots. But those slots cannot usually be filled by other abstract constructions (except for recursive slots such as embedded clauses). Rather, for example a ditransitive construction requires specific types of 'concrete constructions' to fill the roles of the verb and the objects, and those are words (see also Boas (2008a,b)). So implicitly, the lexicon is still split into a number of populations from which a slot needs to select. In a graph-based model, however, concrete and abstract constructions equally may be defined through groups of nodes and/or edges, meaning that this framework allows for their formal modeling as a population of the same space; the same space even with units smaller than words and larger than clauses. This has not yet been done to my knowledge, but with the growing interest for graphs a graph-based implementation of usage-based syntax is likely to appear sooner or later.[8]

Thus, graph-based models hold a promising future for a formal model of usage-based lexicosyntax, because they encompass the potential of a true unification of syntax and lexis through the definition of all relations as subgraphs, and the explicit modeling of the presumed inheritance hierarchy and networking between the instantiations of a construction both within and across levels of abstractness. Until now, however, inheritance hierarchies have only been modeled exemplarily for individual phenomena or lexemes (Zeldes, 2013a; Fried and Östman, 2005); and a view of grammar as the epiphenomenon of lexical subgraphs has been theoretically claimed (Croft, 2001; Hunston, 2012; Hoey, 2012), but has not been implemented or fully formulated in an exact model. Some remarks on potential starting points for future developments of the model here will be made in section 7.2.

Another aspect that makes graphs interesting for the modeling of small- to medium-sized language data is that they are not inferential, and their metrics are not inferential either. Instead, graphs represent elements and relationships between those elements without projecting to presumed external populations. This is helpful where the demarcation of such populations is difficult to achieve, either ontologically, or due to sparse data. It is particularly promising in research concerned with the quantification of lexical data, where with productivity and stark context dependency, it is difficult and perhaps impossible to define the function that projects from the sample to the population. Along with their much higher density of encoded information, they may thus provide the potential for the insightful quantification of limited data. This is of particular relevance in subfields where

[8] In fact, with Hunston's pattern grammar (Hunston, 2012) and Hoey's lexical priming approach (Hoey, 2012), two models already exist that put words at the center of grammar and imply syntax as an emergent or even an epiphenomenon from word co-selection. Both can be translated into a unified graph-based lexicosyntactic model, if syntactic relationships are not understood as re-coselections in in real-time usage, but as somewhat persistent as subgraph edges.

data is naturally limited, such as less-documented languages or historical linguistics, or where data collection is resource-intensive and a lower threshold of the data amounts necessary for quantitative analysis would be of tangible value in research planning.

Out of all linguistic fields, graphs have perhaps played the largest role in semantics, particularly so in lexical semantics in the proximity of cognitive science and digital lexicography. Word and conceptual representations are often modeled as networks, most prominently in WordNet (Fellbaum, 2010) and FrameNet (Johnson et al., 2003), and integrations of both like SemLink (De Lacalle et al., 2014), and those and similar networks are used for the analysis and assessment of the mental lexicon.[9] Some of that work has also successfully mapped aspects of graph structure to linguistic concepts, such as network growth to aspects of language development and performance (Beckage et al., 2010; Kenett et al., 2016) or number of neighboring nodes and cluster coefficients to response time in production (Chan and Vitevitch, 2010), see De Deyne et al. (2016) for an overview.

5.1.2. Summary

This section has presented graphs as a relationship-centered data and knowledge structure that is widely used for data storage, retrieval and analysis in a number of neighboring fields, but not so much in core-linguistics. Graphs were presented to offer modeling advantages in that they are capable of storing and retrieving data of different kinds without requiring a uniquely modeled hierarchy, and that they allow for contradictions and discontinuities, which is favorable for text data in many ways. For the purposes of this work, however, it is the relationship-centeredness that is of particular interest, because it allows for a consideration of the whole distribution of a corpus in a single metric, as will be discussed further below, and it adds a layer of information that can be harvested in the analysis, which is of particular relevance in small to mid-sized data. The rest of this chapter will be used to model a graph-based representation of Kobalt.

5.2. The model

This section introduces the graph-based model of Kobalt that will be used for the measurement of coselectional constraint. In the spirit of "all models are wrong, but some are useful" (George Box), it is first necessary to define its scope of application:
The purpose of the graph-based model in this thesis is to operationalize the question of

- how coselectional constraint across verbs and nouns can be measured in corpora;
- and, due to limitations in time and resources, how this can be done largely automatically.

The model here is *not* a fully formulated model of coselectional constraint in all its potential aspects; and it is *decidedly not* intended to represent a model of syntax, lexicosyntax in general, or lexicosyntax at the lexicosyntactic-semantic interface.

Rather, it is a first approximation of a linguistically not yet well-understood phenomenon (co-selectional constraint or coselectional preference) as measured in a linguistically not yet well-understood type of modeling framework (graph-based modeling). Thus, the model

[9] It should be noted that WordNet, FrameNet, and SemLink depending on the context are often more likely to be seen as databases for NLP and computational linguistics, while other literature views them as representative of the lexicon of a given language.

provided is relatively simplistic. Although it is intended to reflect relevant linguitic distinctions, any distinctions that are not relatively clear-cut and those that require manual annotation and/or the formulation of another model of the phenomenon will be left out at this stage, despite their being some of the most interesting aspects of language. This is not to say these distinctions are irrelevant to the phenomenon, but is a partially pragmatically and partially methodologically driven decision.[10]

Coselectional constraint as a structural property of language, while recognized in the *idiom principle* and Pawley and Syder's *two puzzles for the language learner* (Pawley and Syder, 1983), is generally understudied regardings its specifics. Existing studies largely set out to categorize collocations and make them available for language learners, but are undertheorized with respect to properties of coselectional constraint as such. Thus, insight into what needs to be considered in the operationalization can be drawn mainly indirectly from studies of constructions/collostructions, semantic constraints, and productivity (see sections 2.1.2 and 2.1.3). To summarize some linguistic and pragmatic constraints that should be reflected in the model:

1. Collocation studies vary enormously in which coselections they consider. There exists no classification that would provide an analytical framework for the whole network of coselections in a corpus. Thus, collocations of the verb + object type will be considered without further differentiation by degree of idiomaticity, flexibility, non-compositionality, or other aspects discussed somewhat frequently in regards to the study of collocation. This is not to suggest that further differentiation (by degree of flexibility, idiomaticity, specificity, verb semantics, verb class, etc.) are irrelevant to the measurement of coselectional constraint. However, any such differentiation would require the development of a model theoreof, which can unfortunately not be provided within the scope of this thesis.

2. What is clear is that a positional model that defines coselection collocationally (through the co-occurrence of two words within a set token window) invites randomness and uncertainty into the model in German. Consider the following example: *Sie hörte in ihrer Wohnung oft laut Musik. Sonntags las sie aber lieber ein gutes Buch.* ('She often listened to loud music in her apartment. On Sundays, however, she prefered reading a good book'). Here, 'to read' and 'music' are positionally closer than 'to read' and 'book' and 'to listen' and 'music'. In addition, the verb moves to the sentence-final position in subclauses, *weil sie in ihrer Wohnung gern laut Musik hört* ('because she likes to listen to loud music in her apartment'), where the order of verb and object are inversed compared to first example, and verb and object are now adjacent. A simple n-gram approach would be unable to discern between those cases. Thus, a syntactic model is required. The model – agnostically – chosen here is a simplified version of Foth's dependency grammar (Foth, 2006). There is the argument that a dependency-based approach is also structurally better suited to capture coselectional constraint, because it is a highly lexically oriented syntax model. However, since differences between various syntactic approaches and their implications for the lexico-syntactic interface and coselectional preferences cannot be discussed here, suffice it to say for now that a dependency grammar-based model

[10] Methodologically, because if a re- or a first modeling of a phenomenon is required, the resulting model would be uncertain since it would be modeled on previously seen data and require replication first; and would still require a number of forced choices during annotation, thus skewing the data in ways that are impossible to control for and difficult to estimate regarding their strength or direction.

marks a good pragmatic choice (it is easy to parse and process).

3. Verbs in different senses, and homographs of all categories with distinct meanings, will have different coselectional constraints. This is implied in the fact that collocational profiles are used in verb sense disambiguation (Nastase, 2008; Veronis and Ide, 1990; El Maarouf et al., 2014). However, a problematic aspect here is that it is not entirely clear how to define the boundary of one word sense from another (semantically? By argument structure? Syntactically? Contextually?) – this will be further discussed in section 5.2.1.1; and in the context of usage-based lexicosyntax, one of the pillars of argumentation is the concept of entrenchment, i.e. the strengthening of a connection between forms that are used together. In entrenchment, however, a phonetic or graphematic form should be entrenched with the coselected object *regardless of which sense it appears in.* Thus, there is uncertainty to both the process of disambiguation (how to tell one sense from another) and its place in usage-based modeling. In addition, unless done automatically through algorithms that use collocational profiles for disambiguation – which would make a circular argument – word sense disambiguation requires manual annotation, for which the resources of this project do not suffice.

4. Plank (1984) also suggests that, due to their higher semantic specificity, German prefix and particle verbs have higher coselectional constraints than simplex verbs. Clearly, semantically light verbs have the lowest degree of coselectional constraint (nearly anything can be 'had' or 'given'), while support verbs in support verb constructions (*Funktionsverbgefüge*) have the highest specialized constraints (*role* in *to play a role* cannot be exchanged without losing the verb sense of 'to be important'). Thus, discernability between verb classes seems relevant to the model.

5. Although verb-argument structures are not entirely defined by semantic similarity and a notable degree of idiosyncrasy exists, the idea of semantically motivated cases generally prevails in syntax and lexicosyntax.[11] This suggests that a division by cases is generally a good idea for the study of coselectional constraint. German has four cases (nominative, genitive, dative, accusative), all of which can be construed as subject or object types respectively. Genitive objects have become rare and barely occur in the data, while dative objects exist, but are still much less frequent compared to accusative objects in the data. Other types of objects in German, that are not defined by cases, are prepositional objects, infinitival and clause complements, and predicates.

6. Similarly, from a semantic perspective, both frame- and feature-based semantics stipulate principles similar to the *Semantic Coherence Principle* in Goldberg's Construction Grammar: That slot fillers must have features that agree with the verb semantics in question. For example, Plank (1984) argues that verbs may have very narrow coselectional criteria, such as the verb *meow*, that allows only for meowing subjects (this will be further discussed in section 5.2.1.3). As Plank (1984) suggests and Zeldes (2012) shows empirically, different object slots behave differently regard-

[11] First formulated in modern linguistics in Fillmore's Case Grammar (Fillmore, 1968, 1977), but also in agreement with Cognitive Grammar (Langacker, 1987), and incorporated in the idea of semantically meaningful construction slots (*abstract constructions* in Construction Grammar, see Croft (2001) in particular).

ing the acceptance of a many and/or new slot filler lexemes. This strenghtens the case for maintaining a differentiation between argument slots.

7. In addition, verbs and objects generally form more coherent and interdependent units than verbs with subjects (consider *Alma plays soccer* vs. *Alma plays*; and *Alma plays soccer* vs. *James plays soccer*). This is a sensitive aspect in two ways: Firstly, a differentiation between passive and active voice is required to tell semantic subjects from semantic subjects for verbs, i.e. for the sentence *The High Court made a decision* and *The decision was made by the High Court* to count *decision* and *High Court* semantically as accusative object vs. subject coselections respectively in both cases. This is particularly relevant since Plank (1984) suggests that subject slots are the least coselectionally constrained for most verbs.

8. The second aspect where this becomes relevant is the issue of (un-)accusativity and (un-)ergativity (Levin et al., 1995; Kuno and Takami, 2004). Unaccusativity has been suggested as a property of verbs realizing a syntactic subject on the surface which in fact represents an underlying object, i.e. a semantic patient: *The cup broke.* If the agent is realized, it fills the subject slot: *The boy broke the cup.* If it is not realized, the object moves into the subject slot, resulting in an unergative or non-volitional reading – consider in contrast: *He taught chemistry* vs. *#Chemistry taught.* With it comes the infelicity of impersonal passivization: **It was broken by the cup* vs. *It is taught (in schools) that (...).*

   For the study of coselectional constraint, unaccusativity may play a role, specifically where object slots are analyzed separately, because subjects may be semantic objects depending on the realization of the accusative object. At the same time, it has been shown by Kuno and Takami (2004) that the acceptability of accusative or ergative readings of unaccusative and unergative verbs depends on many factors and cannot be derived from introspection (i.e. without context) or from limited data. Thus, what counts as an unaccusative verb cannot easily be read from a word list, but requires a manual case-by-case annotation, which cannot be realized within the scope of this thesis. For a further development of a model of coselectional constraint, on the other hand, it does appear worthwhile considering unaccusative readings in the model, where it might even be the case that acceptability of one over the other is sensitive to coselectional constraints.

9. Some studies suggest that coselectional constraints may underly morpho-phonotactic principles (Ambridge et al., 2012, 2014). While interesting for future research, it is as of yet unclear what exactly this entails, and it would require a morpho-phonotactic model of German, which cannot be developed within the scope of this work.

In summary, little can be said about the specific linguistic requirements and constraints of the model aside from

- the necessity of a syntactic model in order to account for German word order movement and flexibility;
- a likely benefit from an object type distinction, specifically a subject vs. object distinction;
- a likely benefit from verb class annotations;

- a likely benefit from several aspects that are modeled to a degree, but require manual annotation (word sense disambiguation, fine-grained semantics, unaccusativity).
- a likely benefit of several aspects that are not sufficiently modeled in linguistics yet (role of morpho-phonotactics, classification of coselections);

As it was mentioned earlier, due to limitations in time and resources, as well as for methodological reasons, further manual annotation beyond the annotation of target hypotheses, the correction of dependency parses, and the annotation of verb classes cannot be realized in this work.

The focus on verb + object coselections in this thesis is for two reasons: Firstly, they are syntactically interesting and underly development in several ways (for example, the occurrence of support verb constructions vs. light verb constructions, which are syntactically similar but semantically and lexicosyntactically different; the higher occurrence of prepositional objects in more advanced learners). Secondly, verb + object coselections are abundant in learner language at all levels of acquisition. However, the model in principle provides also the option of analyzing other kinds of coselections. For examples and some problematic aspects of functional words in this respect, see section 5.2.1.3.

Since this thesis provides only a first approximation to the measurement of coselectional constraint as a structural property, not all existing annotations are harvested for measurement in this first step. Rather, object type distinctions are used for a differentiation between levels of specificity (see section 5.2.1.3); and for suggestions for future research in chapter 7.2.2. Verb class annotations are not used in the computations here, although they were used in the statistical analysis in chapter 4.

At all points in the model and the computation, where the decision was made to leave out a presumed distinction, this was done in a way that keeps results on the conservative side. Thus, results are to be understood as a first, minimal approximation and more intricate patterns are likely to be found where further distinctions can be included. Further and more fine-grained annotations can be integrated with the existing data at any time. Object type and verb class annotations are preserved in the graph model and available for analysis immediately.

5.2.1. Specifics of the model

The model of coselection here is not positional, but (hierarchically) syntactic based on a slightly adapted version of Foth's dependency grammar of German (Foth (2006), all changes are documented in this section and section 3.2.2). While there may be some linguistic arguments for considering positional collocations (n-grams) as coselectional in nature, such as a phonetic entrenchment or positional preference for subject and verbs in adjacency, a deeper structural analysis is favorable for an understanding of coselection as a lexicosyntactic rather than a purely lexical phenomenon. Also, German as a language with a somewhat flexible word order is generally less accurately captured in surface adjacency models.

Dependency grammar goes back to the work of Lucien Tesnière (Tesnière (1965), Welke (2011) for a summary with German examples, and Ágel and Fischer (2010) for a more recent overview of valency and dependency grammar approaches) and describes the grammar of natural languages as a series of dependencies between words: Each word depends on another word, except for the finite verb of a sentence that marks the root node. Dependencies can either exist as valency patterns (mostly of verbs, but also some adjectives

such as '[x years] old') or mark adjuncts. Dependencies are labeled, where in Tesnières work there are only few types of dependencies, while Foth distinguishes between a large number of syntactic classes. Valency and dependency approaches have recently played a larger role in usage-based linguistics (Herbst, 2014b; Herbst and Uhrig, 2009; Faulhaber, 2011; Engelberg et al., 2015).

Argument structure is modeled as an abstract level construction in Goldberg's CxG (Goldberg, 2006, 3.4, 4.2) and valence is one of the features that define a verb construction in Sag's sign-based CxG (Sag, 2012, 79ff; 85ff.). Even on a lexically specified level, an interface between valency or dependency grammar and construction grammar appraoches can easily be created through the incorporation of prototype theory, since both valency and construction grammar work with slots or openings that require fillers and the number, type, and perhaps also the order of slots is modeled as meaningful in both valency and constructional approaches.[12]

Dependency grammar in particular has also been embraced by computational linguistics.[13] This is due to its flat structure, where no intermediate layers or higher order phrase structure as in constituency trees is required, movement and with it traces and empty categories can be avoided, crossing edges are unproblematic, and each word can be assigned dependency directly to another word,[14] which is helpful in terms of labeling and parsing because it lowers the overall complexity compared to the number of possible trees in a sentence.

However, since each word can depend only on one other word, coordination is notoriously problematic in dependency grammar. For example, a subject shared by two verbs as in *They like to read and listen to music*, *they* could only be assigned as the subject to either read or listen to music. Some other complications arise not from the nature of dependency grammar as such, but from choices made by the concrete model: Foth's annotation schema follows much of syntactic theory in allowing for only one filler for each slot type. Thus, in the sentence *Frühere Generationen geben uns eine Plattform, eine Basis, einen Ausgangspunkt* ('Earlier generations give us a platform, a base, a starting point', BY_081), the three accusative objects are somewhat artificially analyzed as a single accusative object, upon which the other accusatives depend as coordinated words without further specification of their syntactic embedding (fig. 5.2).

All coordinated words in Foth's original model receive either the label KON or CJ depending on whether a coordinating or comparative conjunction (*als, wie*, 'as, like') is used, and whether they close the list of coordinated items or not (CJ is list-final). These labels are the same for all linguistic categories, i.e. no differentiation between object types, cases, or parts of speech is made: "Coordinations are modeled as a right-branching chain of words. Each word carries the label KON, except for the final word, which is subordinated to the conjunction and is labeled CJ", (Foth, 2006, 22, my translation).[15]

[12] Unifying construction grammar with phrase structure and transformational grammar approaches requires modifications to both models, but some work into a synthesis with HPSG exists, see Müller (2013b); Richter and Sailer (2009).

[13] See Rehbein (2010) for an overview. With the universal dependency parser (Nivre et al., 2016; Nivre, 2015), the influence of DG in computational linguistics has further grown, and seeing that it is already being used by a number of corpus projects and will likely further increase in relevance in the future, there is a high chance that DG will play a larger role in future core-linguistics, too.

[14] This is not to say that dependency grammar does not or cannot imply a grammar of word groups, too, as Engel (1996, 54) notes.

[15] The German original reads: "Beiordnungen werden als eine rechtsverzweigende Kette von Worten modelliert. Dabei trägt jedes Wort das Label KON, bis auf das letzte Wort, das unter der Konjunktion steht und mit CJ bezeichnet wird".

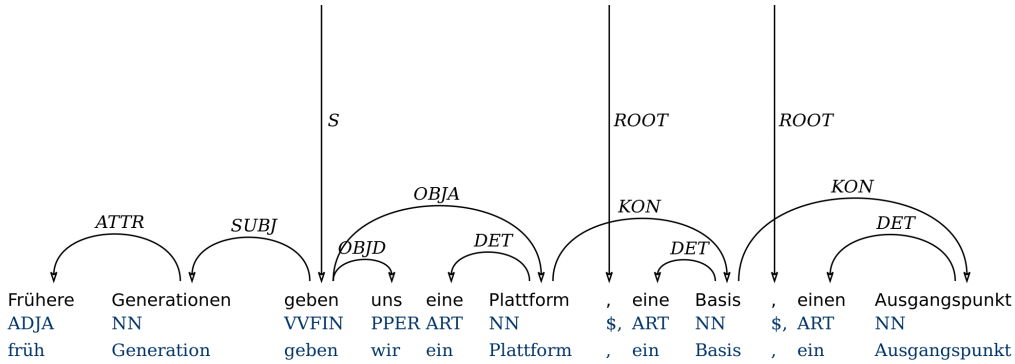


Figure 5.2.: Dependency parse with several realized accusative objects cf. the original model in Foth (2006): Accusative object realizations are chained as coordinations without specification of type and depend on the prior element of the coordinated chain. All dependency visualizations are done with `https://arborator.ilpga.fr/q.cgi`.

This is linguistically unenlightening and annoying in data parsing, because the distinction between different object types and even parts of speech is lost: Consider 'my <- one <- and <- only', where in dependency labels, *one* is a determiner, while *only* is an attribute (both to a noun ellipsis); While in Foth's model, *only* would be labeled CJ and could not be told apart from the accusative objects above without including the part-of-speech annotation in the search. In the case of object types, however, part-of-speech tags would not be helpful for distinguishing between object types either, since OBJA, OBJD, OBJG all are nouns. Since the parses were manually corrected anyway, I have chosen to label all objects and subjects as dependencies of the verb directly (see fig.5.3).[16]

Another shortcoming of the theory in the context of the study of coselectional constraint is that the verb group is split up into dependencies in a way that lexical verbs depend on auxiliaries. This is consistent with the idea of the finite verb as the root of the sentence and all other words as its dependents, but it also implies a split analysis of subject and object dependencies, where subjects are assigned dependency of the finite verb, while objects depend on the lexical verb. Those two are the same verb for lexical verbs, but not in modal and auxiliary constructions, i.e. all analytical tense, aspect, mode (TAM) constructions in German, such as *Wir haben Käse zum Frühstück gegessen* ('We have eaten cheese for breakfast')) or *Er muss heute die Katze füttern*, ('He needs to feed the cat

[16]This annotation schema and Foth's approach can be transformed into one another, but the data in the repository (`10.5281/zenodo.3584091`) is labeled as described here. There are ways to reconstruct object labels from Foth's analysis, too. These require a specialized search for each case, for example: *For all CJ, find the words they depend on, if those are objects of a certain kind, add the word to the list of dependents of the verb those depend on, otherwise repeat recursively.* However, this is not only slightly inconvenient with the CONLL format in which the data exists, but also seems unnecessary, because the conflation into a coordinated subclass appears theoretically unmotivated in the context of a usage-based application of dependency grammar: The assignment of KON and CJ exists in order to avoid the possibility of multiple argument realization. From a usage-based perspective, there is no necessity to avoid multiple argument realization though, and it is strange to suggest that the order of elements in a larger coordinated set should lead to different syntactic analyses: In the example here, *eine Basis, eine Plattform, einen Ausgangspunkt* would yield a different syntactic analysis than *eine Basis, einen Ausgangspunkt, eine Plattform* or other permutations, where the dependency of words depends on which other word one is most proximal – which is precisely what is attempted to avoid in the analysis here.

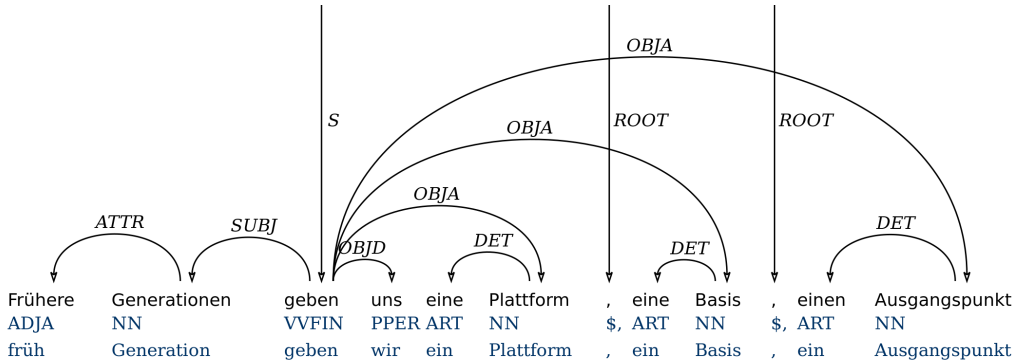


Figure 5.3.: New dependency parse with several realized accusative objects according to the model developed here: All accusative objects depend on the verb directly and are labeled as OBJA.

today')). Fig. 5.4 shows an example from Kobalt with the original parsing: *Heute können junge Frauen einen solchen Mann nicht finden, dem sie ganz vertrauen können*, 'Today, young women cannot find such a man whom they can fully trust' (BY_026).

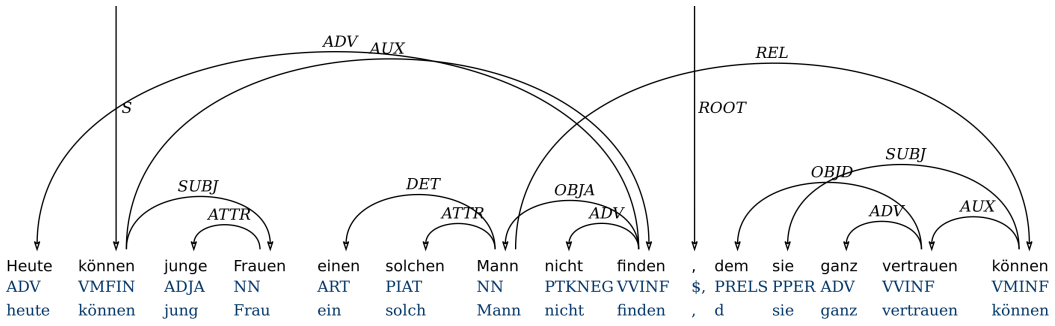


Figure 5.4.: Dependency parse with two modal constructions cf. the original model in Foth (2006): Subjects are assigned dependency of the finite verb, here the modal verb *können*, 'to be able to, can'. Extracting the coselections from this parse to a [V SUBJ OBJA OBJD]-schema results in [können Frau ∅ ∅] and [finden ∅ Mann ∅]; and [können sie ∅ ∅] and [vertrauen ∅ ∅ d(em)] respectively.

The model has been adjusted so that subjects were labelled to also depend on the lexical verb, see fig. 5.5, and has two effects:

- Agreement information between subject and finite verb is lost.[17] This is irrelevant to the present analysis, because no claims regarding varying inflectional forms in coselectional constraint have been made. However, for a contrastive analysis of chunks and coselectional preferences, this might require reconsideration.
- The analysis implies a lexical or semantic focus, where the coselectional preferences of a verb are tied to the lemma *per se* rather than the realized form including syntactic context. It is, however, possible, that there is an interaction between modal verbs or analytical TAM constructions and coselectional preferences. In a more form-focussed model, such differences would require reconsideration. However, this would

[17]The analysis here is based on a lemmatized layer, but a graph from tokens can easily be reconstructed.

also imply that *Heute finden Frauen einen solchen Mann nicht*, 'Today, women do not find such a man' and *Heute können Frauen einen solchen Mann nicht finden* 'Today, women cannot find such a man' are *coselectionally* different, i.e. do not instantiate the same coselectional choice. This again concerns the theoretical and empirical identification of boundaries between chunks and coselections and requires more theoretical modeling in future extensions of the model.

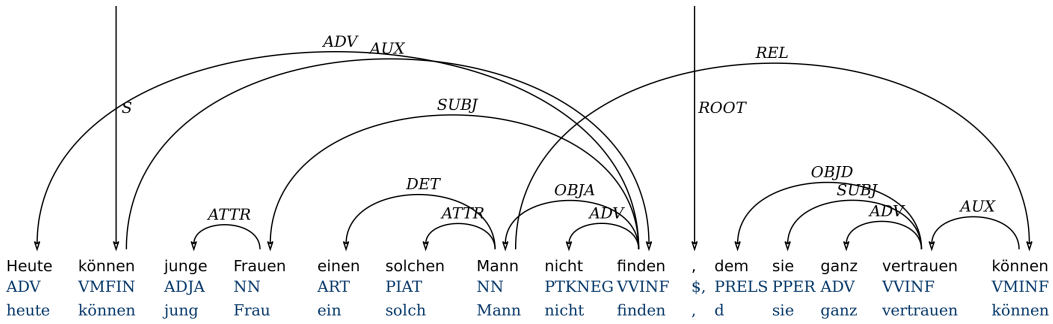


Figure 5.5.: New dependency parse with two modal constructions according to the model developed here: Subjects are assigned dependency of the lexical verb, here the modal verbs *finden*, 'to find', and *vertrauen*, 'to trust'. Extracting the coselections from this parse to a [V SUBJ OBJA OBJD]-schema results in [finden Frau Mann ∅]; and [vertrauen sie ∅ d(em)] respectively, thus maintaining all coselections of the lexical verb.

With the corrected dependencies as a basis, the model requires the following specifications:

1. What is a node?
2. What is an edge?[18]
3. Do nodes and edges have any additional properties, such as
   - edge weight?
   - edge labels?
   - node size?
   - node properties or labels?

The following sections document the choices in some linguistic detail. A formal definition of the graph-based model can be found in appendix A.1.

5.2.1.1. Nodes

Nodes are modeled as lexemes without word sense disambiguation. As has been mentioned earlier, this is an obvious oversimplification of the underlying linguistics in the sense that distinct behavior is to be expected from diverging word senses, certainly in the case of

[18] There are several papers discussing this from a theoretical or applied perspective, see for example Woods (1975); Arias-Trejo and Plunkett (2013); Koolen and Kamps (2009). The model here, however, is limited to a pragmatic choice of what can reasonably be extracted from the corpus and considered a coselection.

homographs, but also to a lesser extent in polysemous words. However, the discussion of what constitutes a word sense has a long history in linguistics and annotating word senses is complicated and resource-intensive. The model here is therefore based on the corrected TreeTagger lemma tags (Schmid, 1995). This is particularly unsatisfying for the verb domain, where functional and lexical senses of verbs are conflated in a way that sometimes creates cycles in the graph (*Sie haben etwas gehabt* ('They had had something'), *Sie sind dort gewesen* ('they had been there', where the auxilliary is realized with *sein* ('to be') in German)). Morphosyntactic categories can in fact be distinguished in the data since they are annotated accordingly (see section 3.2.2), but for the graph model it appeared as an artificial division for two reasons:

- Firstly, the verbs most affected by this are *haben* and *sein* ('to have', 'to be'), where auxiliary uses are filtered out on the higher levels of specificity anyway (see next section);
- Secondly, the other category of verbs affected by this are those that appear as either lexical or support and construction building verbs.

Both are, by definition, semantically idiosyncratic and have fuzzy category boundaries. For example, the verb *zeigen* ('to show, to point') has a ditransitive use that is semantically straightforward:

(1) Eine dieserartige Unselbstständigkeit kann uns zeigen, dass die Möglichkeit
one such lack_of_independence can us show that the possibility
hoch ist, dass die Menschen in schwierigen Situationen immer die Hilfe von
high is that the humans in difficult situations always the help of
anderen brauchen.
others need.
'Such a lack of independence can show us that there is a high chance that people in difficult situations will always need the help of others' *BY_069*

However, there are other uses that are less clearly defined, as in the reflexive construction *sich zeigen* ('become evident, clear' or 'can be seen', literatlly 'show itself'). This can be fully passive and unintentional, as in:

(2) Wie sich zeigt, ist das Leben für einige unserer Generation tatsächlich
How itself shows is the life for some our.gen generation actually
einfacher.
simpler
'As can be seen, life for some in our generation is actually simpler' *DEU_018*

Or it can have an aspect of intentional expression:

(3) Es zeigt sich in der materiellen und auch der mentalen Welt.
It shows itself in the material and also the mental world
'This becomes evident/is expressed in the material and the mental world' *CMN_021*

This is not bound to the syntax:

(4) Sie haben mehr Freiheit und können was tun, um ihren eigenen Wert
They have more freedom and can what do in_order their own worth

zu zeigen.
to show
'They have more freedom and can take action to show their own worth' *CH_058*

In this case, a syntactic distinction would still conflate word senses, while a semantic distinction would still contain different syntactic structures, therefore breaking through lexicosyntactic constraints of the other constructions included; and a clear semantic distinction, for example with respect to the question of intentionality/expression in the example above, is not always possible to make. Moreover, the analyses here are usage-based in nature, and the level of differentiation between word senses for homographs or homonyms is not always clear or clarifiable in learners.

In addition, there is the problem of entrenchment, as mentioned previously in this section and in section 2.2. If entrenchment, i.e. the progressive strengthening of a connection between two items through their co-occurrence, is in fact bound to a phonetic or a graphematic form, homographs or homonyms should correctly be treated as one form. If, however, a semiotic model is implied, then there is not only a problem of distinguishing between word senses, but also the problem of semantic clustering by semiotic similarity.

Against the background of this uncertainty I have therefore decided to model conservatively with lexeme homographs represented as a single node regardless of their potential or obvious ambiguity or polysemy, accepting the possible overwriting of constraints in some cases. This means that results will tend to underestimate levels of lexicosyntactic constraint or structural sophistication of the graph, which is a strategy I will follow in some other respects, too (see next section and chapter 6 for more details).

Nodes in this model aside from the identifying label have five properties. Three of those are represented in the final graph data (.json-files) and partially in the visualizations:

- *subcorpus frequency (sc_freq)*: The frequency within the given subcorpus;
- *document count (doc_count)*: The number of documents a lexeme appears in in a given subcorpus;
- *category (cat)*: The morphosyntactic verb category (such as modal, particle, prefix, simple lexical verb, details can be found in chapter 3.2.2.

Since homographs are modeled as a single node, while some verbs can have more than one function (*haben* 'have' can be a lexical verb, an auxiliary, or part of a modal infinitive construction), there is a 'mixed' category label in the final version of the graphs. The original categorization is preserved in the data and can be accessed if necessary. Two more properties are required for filtering subgraphs by specificity, *part of speech (pos)*, and *passive (pass)*.

In the following visualizations, node size represents *document count*. In the computations that follow, however, the overall frequency in a subcorpus is considered implicitly because the measure applied is based on the distribution of outgoing or incoming edges which in sum translate to word frequency.

5.2.1.2. Edges

Edges represent dependency, where dependency labels (SUBJ, OBJA, etc.) are encoded as edge labels. Edge direction is encoded through the order of source and target nodes and

reflects dependency (dependent = target). Lexicosyntactically, there is a case for modeling the graph as undirected for at least two reasons:

- Dependency goes one way in most cases (verbs govern nouns), but there are some exceptions with participles used as deverbal adjectives, as in *entwickelte Länder* ('developed countries'), where *entwickelte* is an attribute to *'Länder'*, while in the predicative version (which is semantically similar, though not identical) *Diese Länder sind entwickelt* ('These countries are developed'), the noun depends on the verb. In a directed graph, this would double the edges for those two lexemes. It is not implausible to assume that those two differ in usage and meaning, but splitting the co-selection of two identical lexemes, *entwickeln* and *Länder*, into two distinct categories based on their occurrence in a predicative vs. attributive slot seems redundant and overly complex.

- Association, priming, entrenchment, and entailment are, unlike dependency, not directionally specified. A semantically chosen object may prime or entail a certain verb, and sometimes this is even grammaticalized or lexicalized as in the case of support verb constructions (*Funktionsverbgefüge*).

What makes it necessary to formally define edges as directed, however, are the edge labels, which are implicitly directed (OBJA conceptually has a verb source and a noun target) and would render the model inconsistent if used in an undirected graph.

Edge weight corresponds to the absolute number of co-occurrence of source and target nodes in the subcorpus. Edges have one label *dependency (dep)* representing the dependency label (such as OBJA, OBJD, etc.).

5.2.1.3. Specificity of the graph

The question of graph specificity refers to which lexemes should be included in the graph relative to a continuum of including all lexemes in all syntactic positions to including only the categories relevant to the analysis, requiring a definition of what those are.

When we look into verb argument structures, what is usually included in analyses regarding coselection are the verbal head and the slots specified by the construction or the subcategorization or valency frame (depending on the theory used). These are typically object or predicative arguments. Some verbs also require arguments that are not grammatical objects in NPs, but rather adverbial or adjectival complements.[19] What constitutes an obligatory argument cannot be uniquely defined for most verbs or verb senses, and non-object complements therefore cannot easily be reliably extracted automatically with high accuracy. Some verbs also govern prepositional objects, where the preposition is fixed in a presumed verb signature, while others show a strong preference for PPs without constraining the chosen preposition.

Thus, a continuum from unspecific to specific exists for potential graph specificities: The least specified, most random graph includes all lexemes, whereas the most specified includes only the verb and its non-subject complements as they are defined by traditional approaches. Since this thesis argues from a usage-based perspective, no claims to the obligatoriness of arguments or predicates are made. Graphs are generated from the parsed data, which are based on target hypotheses 1 (ZH1, Reznicek et al. (2013, 2010)). In

[19] ?*Er wohnt*, 'He lives, resides', where a specification of living conditions or location is obligatory with the verb *wohnen* ('to live, to reside') as opposed to *leben* ('to live') in most cases in German; But *Er wohnt allein*, 'He lives alone', *Er wohnt in Bremen*, 'He lives in Bremen', examples from Müller (2013a, 13).

those, congruence is established in unclear cases. Where a formal categorization as, for example, a dative object is possible, it would be left a dative object even if the verb would not typically allow for a dative object. Predicate and prepositional object status was assigned generously (see chapter 3.2.2) in order to include as many coselections that may be closely tied to the verb as possible. With the lack of theoretical modeling of coselectional constraint, it is hard to tell what might act as an ideal representation for the precise measurement of its development. Thus, the question of graph specificity is a question of intelligent filtering of the data rather than a new model, where more specific graphs provide better ground for analysis, because they are less noisy. Eventually, with a better model of coselectional constraint, adjustments will likely be necessary.

Aside from representing differences in the grammatical model, including certain lexemes in a graph also creates artificially high connectivity: Function words in particular work as a hub, connecting nouns and verbs that are not actually connected in the base text, like in fig. 5.6. The same happens through prepositions in a verb-argument-specific graph that includes the preposition of prepositional objects as can be seen in fig. 5.7. Of course there is always a path from one node to another in any connected component of a graph. However, here, these function words artificially create highly connected communities centered around a single determiner or preposition, which creates noise in the analysis of coselectional constraint with the metric chosen (see next section). It is also not a great representation of the theoretical model, because prepositions and determiners are classes with relatively few members, are ubiquitous in language, and follow a number of syntactic and lexical rules. Thus their occurrence is defined by many more aspects than coselectional preference. However, it is also true that prepositional objects and semi-obligatory prepositional phrases mark very special cases of coselectional constraints and preferences. An intelligent inclusion of these in the graph-based model should therefore be sought out in the future. For a research question centered around coselectional constraints, the

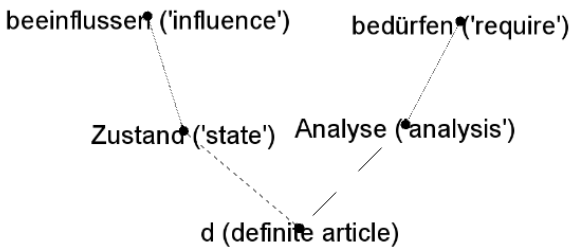


Figure 5.6.: Example of a determiner connecting unconnected verbs and arguments in Kobalt L1, visualization done with gephi (Bastian et al., 2009).

model profits from their exclusion from the analysis and a focus on nouns, pronouns, and words that can act as predicates on the argument side, and verbs on the verb side.

Regarding the linguistic model itself, as has been mentioned earlier, a distinction between objects and subjects is likely beneficial to the analysis, since subjects are treated differently from objects (OBJA, OBJD, OBJG, OBJP) in most of grammar theory and are not even in all cases considered verb arguments. There are several reasons for this. Firstly, they behave differently from other arguments syntactically:

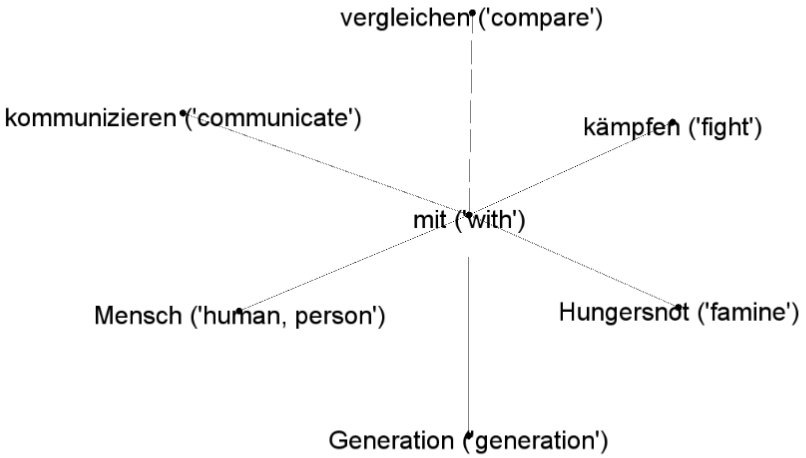


Figure 5.7.: Example of a preposition in a prepositional object connecting unconnected verbs and arguments in Kobalt L1, visualization done with gephi (Bastian et al., 2009)

- They take specified word order positions in many languages (such as VSO vs. SOV etc.), and are often adjacent to the finite verb in unmarked sentences;
- Their morphological form in nominative-accusative languages like German is unchanged by the verb (always in nominative); unlike other complements that are governed by the verb in terms of case, preposition, and sometimes other categories such as the realization or deletion of determiners ('they play a game' – #'they play a football');
- While the number of the object is variable without changes to the verb: *Sie haben den Hund gefüttert* ('They have fed the dog'), vs. *Sie haben die Hunde gefüttert* ('They have fed the dogs), subject and verb agree in number and person (in German);
- They are lost as arguments in passivization, unlike some other object types:
  - *Die Leute sahen die gleichen Bücher und Filme* ('People saw the same books and movies', *CMN_012*) -> *Die gleichen Bücher und Filme wurden [von den Leuten] gesehen* ('The same books and movies were seen [by the people]'), where the prepositional phrase appears as similar to other prepositional adjuncts like [in those times] or even adverbial phrases like [there] or [again and again];
  - But: *Meine Mutter erzählte mir immer ihre unglücklichen Erfahrungen während ihrer Kindheit* ('My mother always told me her unhappy experience during

her childhood', *CH_052*)-> *Ihre unglücklichen Erfahrungen wurden [mir] [von meiner Mutter] immer wieder erzählt*, where [mir] remains in dative;

– *Aufgrund der Armut soll meine Mutter sich nach der Schule immer mit dem Land beschäftigen* ('Due to poverty my mother was supposed to take care of the land after school', *CMN_052*) -> *Aufgrund der Armut sollte sich [*von meiner Mutter] immer [mit dem Land] beschäftigt werden*, where the prepositional object [mit dem Land] remains intact, while the transformed subject is ungrammatical due to the reflexivity of the verb.

In transformational grammars, such as X-bar theory, government and binding, or generative minimalism, as well as lexical functional grammar, subjects are not included in the VP, but attach to the VP as specifiers in an inflectional or tempus phrase (IP/TP, see Müller (2010), chapters 3 and 6 for an overview). This means that theory-internally, mutual government (c-command) exists between the subject and the VP or I' head node (VP including all its elements), but the verb itself has no c-command over the subject, because the verb is positioned lower in the tree. While specified lexical selection is not typically seen as part of those theories anyway (and neither are subject or object types per se), no c-command also means no access to the individual leaves of the tree from higher up after a phrase is formed. Therefore, while a verb-argument *complex* (a complete VP) could be selective in terms of lexically permitted or preferred subjects, the verb itself and the subject have no access to one another in the tree per se, so that even if one wanted to somehow include argument selection preferences into the signature of verbs or potential non-subject arguments, those preferences would not be able to cross over to the merging place of the VP and the subject in the IP. Some theorists go as far as to conclude that for this reason, idioms cannot be both fully lexicalized including the subject, and have an exchangeable object slot (see Müller (2013a), p. 45–52 for a discussion and counter-examples from English and German).

Other grammar theories such as categorial, construction, dependency and head-driven phrase structure grammar (HPSG) do view subjects as part of the subcategorization frame or argument structure of the verb and account for it in the number of valency slots, and, where they exist, semantic signatures (Müller, 2013a, 2010; Boas, 2013; Kay, 2005). In those, subjects and objects can potentially be treated as equally selective where that applies.[20]

The semantic reason for the higher perceived randomness of subject selection vs. object selection – in German at least – is that, while they are syntactically obligatory in most sentences, subjects contribute less to the meaning of the verb complex. Consider for example the act of playing soccer or watching TV in 'Alma likes to play soccer' or 'Ben watches TV'. Taking away the direct object changes the meaning of the verb complex drastically ('Ben watches', 'Alma likes to play'), while the action itself does not change from exchanging the subject.

This is not to say that subjects are not also coselectionally restricted in terms of semantic categories, such as animacy, concreteness, or collectivity (as for example the German verbs *auseinanderklaffen* ('diverge, gape open') or *zusammenkommen* ('gather') that require a plural or collective subject such as 'positions' or 'people', examples from Starke (1974)) or

[20] Out of these, with its phrasal approach, construction grammar is perhaps most likely to present lexically specified selectional constraints, but it does not typically do so. Instead, descriptions in CxG usually focus on the generative aspects of a construction slot and its selection by verb features, incorporating also the idea that verb senses can be told apart by their subcategorization frames or argument structures (rather than the concrete words they appear with), see Kay (2005); Boas (2013); Goldberg (2006, 1995).

semantic coherence, such as the verb *to bark*, that requires a subject lexeme that can be reasonably construed as being capable of barking. Also, most approaches of construction grammar require semantic congruence between the slots and slot fillers of a construction, such as a verb that can, easily or by an analogical stretch, be construed to have ditransitive meaning. Perhaps the most seminal example of this is *He sneezed the napkin off the table* vs. ?*He waited the napkin off the table*, where a CAUSE-MOTION aspect can be somewhat easily attributed to the first verb but not as easily to the second (for semantic restrictions on construction slots, see Goldberg (2006)). Restrictions like these consistute relatively abstract or logical constraints on subject selection. Aside from those there are also more prototypical agents of actions such as cats for scratching or knives as a subject in an instrument role for slicing (examples from Jarvella and Sinnott (1972)). Plank (1984) lists many restrictions of this kind for specific verbs, but argues that there are higher-order syntactic restrictions on direct and even more so on indirect objects, effectively suggesting a continuum of coselectional constrainedness. In an interesting parallel, this continuum takes the same trajectory as what in syntax theory has become known as the obliqueness hierarchy of syntactic activity (Müller, 2010, 2008), that describes the potential syntactic activity of grammatical categories and relates to aspects of case marking and passivization. While some semantic restrictions do therefore apply to subjects, it remains true that the expression of the act of scratching or slicing is less impaired by exchanging the prototypical subject than the action of a verb-object complex is by exchanging or leaving out the object. In some languages, like Mandarin Chinese, this is even further grammaticalized in so-called verb-noun-compounds, where for example the equivalent to the verb 'to run' is 'to run steps' (跑步, pǎobù). This will be discussed in the context of results in section 7.1.1.

Considering all of these constraints, five levels of specificity were extracted for the graph-based analysis of Kobalt:

1. the full graphs, including all lexemes used as a randomness baseline labelled *full graph*;
2. graphs that include verbs with their subject and object arguments and PPs to see if there is much of a difference between OBJP and PP inclusive graphs labelled *pp*;
3. like 2., but without PP labelled *vas_prep*;
4. graphs that include verbs and their subject and objects including the complement of the prepositional object (the noun that is governed by the preposition in OBJP), but not the preposition itself labelled *vas_no_prep*;
5. like 4., but without subject lexemes, where subjects are defined semantically (OBJA in passive voice, SUBJ in active voice), labelled *no_subj*.

For all specificities higher than *full graph*, only verbs that take objects (including object clause and infinitive complement heads, which are verbs themselves), but not auxiliaries or modals in TAM constructions are included in the analysis. To give an exmaple from Kobalt:

(5) Im Jetzt sollen sich die Jugendlichen nicht mehr um Landarbeit,
In.the now should refl.pron the adolescents no more about farm_work,
Essen, Kleidung und so weiter sorgen
food, clothes and so on worry

> 'In the now, adolescents whould no longer worry about farm work, food, clothes and so on' (CH_052)

The following lexemes are included in each level of specificity:

- full graph: all lexemes;
- pp: *sorgen, Jugendlichen, im, jetzt, um, Landarbeit, Essen, Kleidung*;
- vas_prep: *sorgen, Jugendlichen, um, Landarbeit, Essen, Kleidung* (*im jetzt* is deleted because it is not a prepositional object as clearly defined in the verb signature as *sorgen um*);
- vas_no_prep: *sorgen, Jugendlichen, Landarbeit, Essen, Kleidung*;
- no_subj: *sorgen, Landarbeit, Essen, Kleidung*

Obviously, with progressive exclusion of lexemes, the more specific graphs are also progressively smaller, which will be discussed in the following chapter where it is relevant for the interpretation of results. Figs. 5.8 and 5.9 show most of the resulting *vas_no_prep*-graphs for L1 and BEL-115. It is clear at first glance that, while some visible differences may exist, a visual assessment is not going to suffice for a deeper understanding of structural differences. This is why a metric suitable for the comparison of graph structures will be introduced in the next section. Since the full graphs are not printing format-friendly, they are not included in the appendix. Instead, graph visualizations in *.svg*, a scalable vector graphics format that allows for unlimited zooming and a better visible inspection, are made available with the rest of the data and scripts via `zenodo.org`, a long-term data repository dedicated to data sustainability an dopen science (`doi:10.5281/zenodo.3584091`).

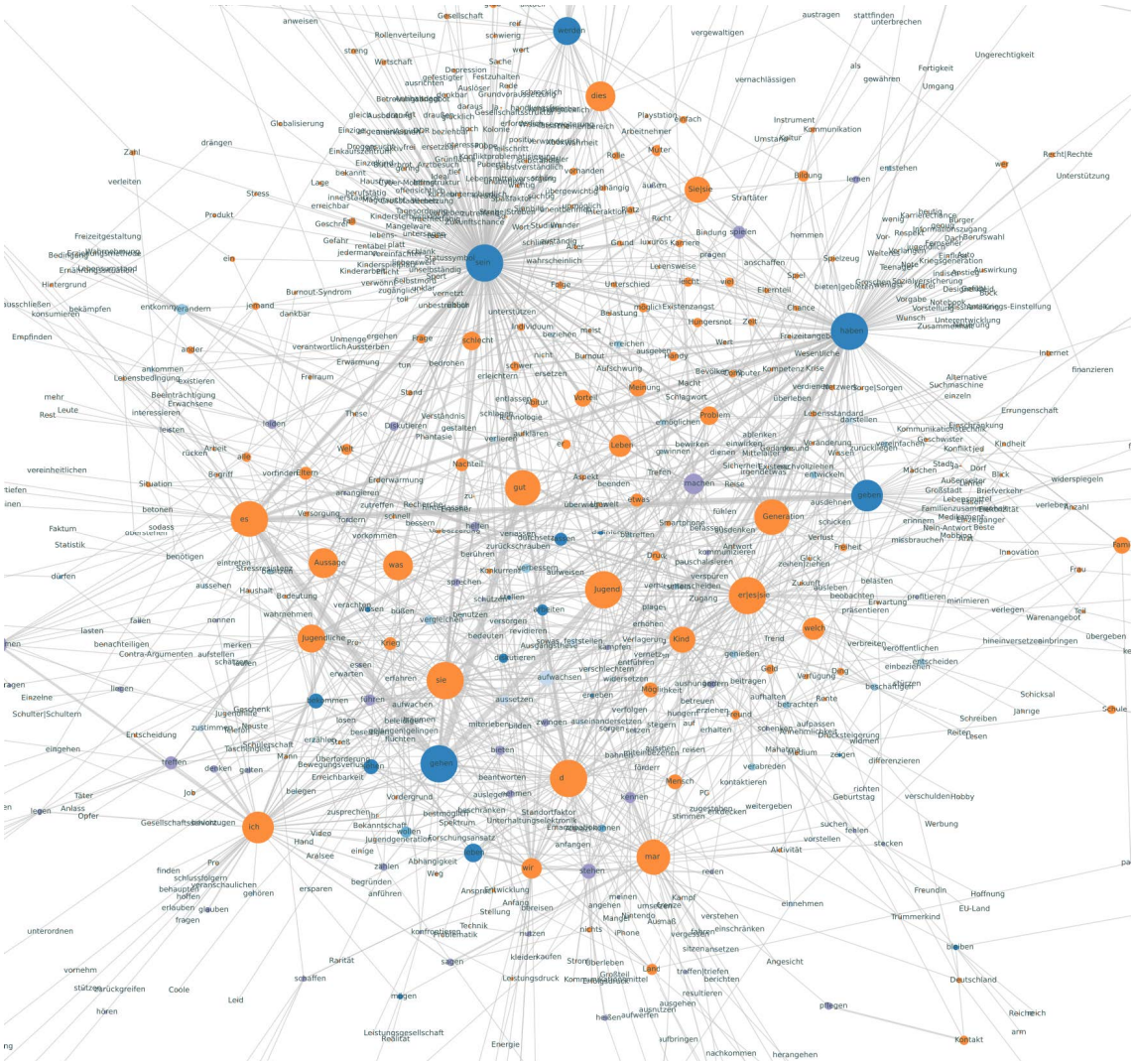
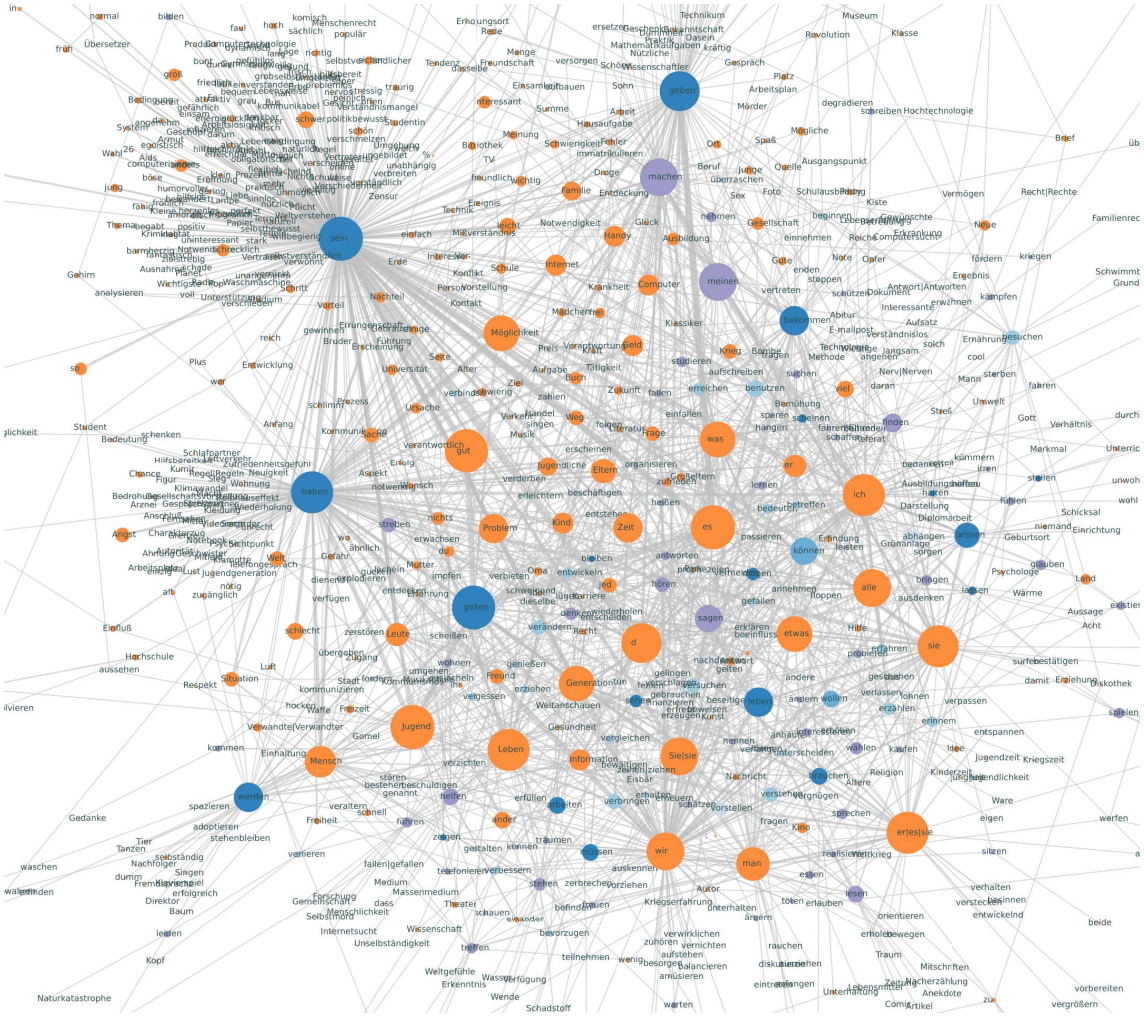


Figure 5.8.: Graph-based verb-argument coselection model of Kobalt L1: *vas_no_prep*. Colors correspond to verbs (blue and violet) vs. arguments (orange), node size to *doc_count* (number of documents in a subcorpus in which the lexeme occurs), and edge width to frequency of co-occurrence. Visualization done with D3.js (Bostock et al., 2011).


sein
haben
werden
machen
geben
bekommen
meinen
gehen
Leben
Jugend
Mensch
Problem
Möglichkeit
Zeit
Kind
Frage
Buch
Arbeit
Schule
Internet
Computer
Familie
Leute
Gesellschaft
Ausbildung
Beruf
Geld
Krieg
Liebe
Angst
Freund
Eltern
Generation
Information
Sie|sie
er|es|sie
wir
man
ich
es
was
alle
sie
d
Gedanke
Freiheit
Natürkatastrophe
Schadstoff
Weltgefühle
Erkenntnis
Wende
Wasserverfügung
Unselbständigkeit
Internetsucht
Wissenschaft
Massenmedium
Medium
Forschung
Gemeinschaft
Menschlichkeit
Selbstmord
Kopf
Baum
Direktor
Nachfolger
Tanzen
Tier
Gehirn
Wahl
Alkohol
Aids
Übersetzer
Hochschule
Einfluß
Gefahr
Arbeitslosigkeit
Arbeitsplatz
Arbeitsplan
Mathematikaufgaben
Hausaufgabe
Mathematikprüfung
Hochtechnologie
Revolution
Museum
Technikum
Klasse
Platz
Mörder
Sex
Foto
Schulausbildung
Vermögen
Recht|Rechte
Familienrecht
Schwimmt
Grund
Ergebnis
Ernährung
Gesuchen
Verhältnis
Gott
Merkmal
Schicksal
Geburtsort
Einrichtung
Psychologe
Land
Wärme
Aussage
Acht
Diskothek
Erziehung
Kriegszeit
Jugendzeit
Kinderzeit
Religion
Ware
Werbung
Werbefernsehen
Werbeagentur
Werbeindustrie
Werbeplakat
Werbeslogan
Werbespot
Werbung

5.3. Graph structure and lexicosyntactic coselection

Based on the graph model presented in the previous section, the next challenge is to operationalize the concept of coselection in quantifiable way, i.e. to find a graph metric which is capable of adequately representing the suspected changes. Interestingly, there are at least two studies that model collocation or co-occurrence as a graph, but then use or suggest using those graphs for extraction of collocations in the traditional way of quantifying lexical association through statistical measures (Brezina et al., 2015; Proisl, 2019). However, it has been shown that a statistical approach is not well suited with the data and the problem at hand. A first graph-specific metric that might instead seem appealing is degree distribution, which counts the number of incoming and outgoing edges per node and translates to the number of collocations, collostructions or combinations (such as n-grams) a word appears in. It thus shows the distribution of the combinatorial power of lexemes in the corpus, where a high degree reflects many coselections or a tight integration of a node into the network, while a low degree shows high selectivity or low frequency. This has been used by Kapustin and Jamsen (2007) in reference to Solé et al. (2010) to show that word co-occurence networks have specific quantitative or structural properties. They find a core vocabulary of a certain size (10^3-10^4 word forms or lexemes respectively in their analyses) that is largely connected and interconnected, and any number of words around those that have very low node or vertex degrees, a so-called *small world effect*.

The two cited studies show this for Russian and English, where Kapustin and Jamsen (2007) suggest their results might a) provide a confirmation of the ergodic hypothesis (sadly, the authors do not further expand on this), and b) constitute a lexical universal. Of course the fact that a large number of nodes have low degrees is somewhat trivial given that in a Zipf-distribution, any larger corpus will consist to a large degree of hapax legomena (words that occur only once in the corpus), which, depending on what an edge is modeled to represent – syntactic, positional, or other kinds of co-occurrences – have a very restricted number of potential edges. Some more remarks on this will be made in section 7.3.2.

Figs. 5.10 – 5.12 (where the last x-axis tick represents the maximum degree for better legibility) show that for the corpora based on the original Kobalt corpus, degree distribution seems to express some of the differences between the no_subj graphs, such as the relatively higher number of hapaxes in L1 and the higher maximum degree in the learner graphs compared to corpus size (CH-130 is smaller than L1 and has an equal highest degree, BEL-130 is 17% larger but has a 30% higher maximum degree and a comparable number of hapaxes) and slight differences in the density of the distribution. However, those numbers are hard to interpret in terms of how different the distributions truly are and what that says about structural similarity or difference.[21] Rather, degree distribution seems to add another zipf-distributed layer, confirming the results in Kapustin and Jamsen (2007) and Solé et al. (2010) mentioned above. Not only lexemes, but also their combinations are power-law-distributed, as they must be, because each word enters an average number of potential coselections defined through the limitations of syntax.

Differences in structure are, however, actually visible in these three graphs in that larger and more independent clusters connecting to a single node that look like bubbles appear in the top right corner (CH) and the right hand side (BEL) of the learner graphs. Similar

[21] Degree distribution plots for vas_no_prep graphs and the lower intermediate learners (OnDaF < 115) are quite similar and can be found in the repository (`10.5281/zenodo.3584091`). The other subcorpora differ in size by a larger degree, making it more difficult to accurately compare degree distributions.

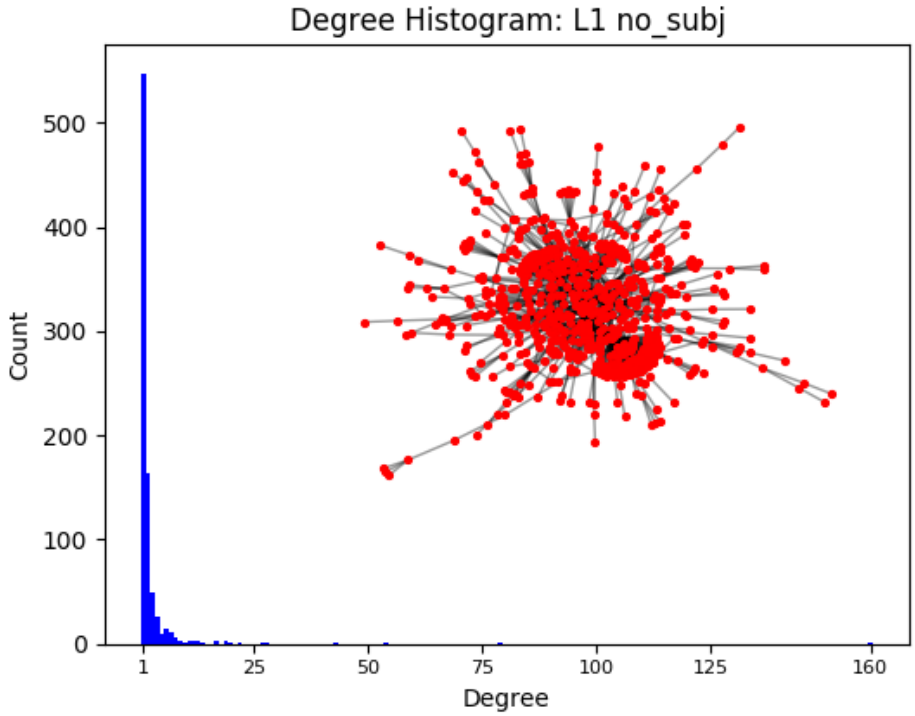


Figure 5.10.: L1 no_subj graph and degree distribution

clusters exist in the L1 graph, too, but they are smaller, less prominent in the graph structure and more embedded.[22]

[22] These visualizations include largest components only. Some of the graphs also contain free-floating nodes that are disconnected from the rest of the graph. This is considered in the computations that follow, but not in the visualization, for technical reasons. A force-directed graph visualization tends to push such disconnected nodes relatively far away from the largest component, making it difficult to adjust the canvas size in a way that the whole graph including its disconnected communities is captured while the individual nodes stay somewhat discernable.

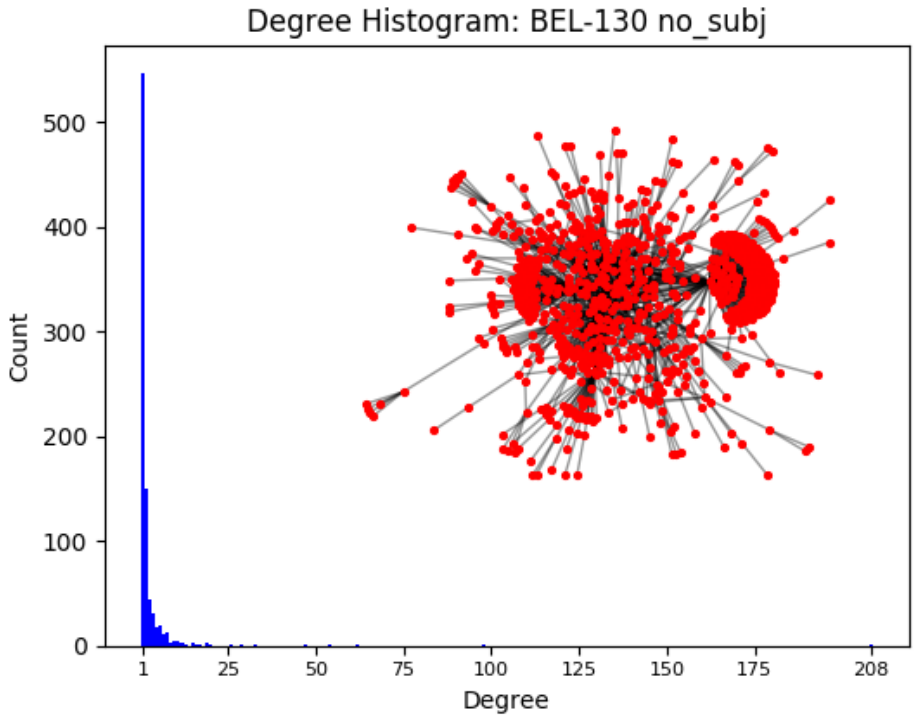


Figure 5.11.: BEL no_subj graph and degree distribution

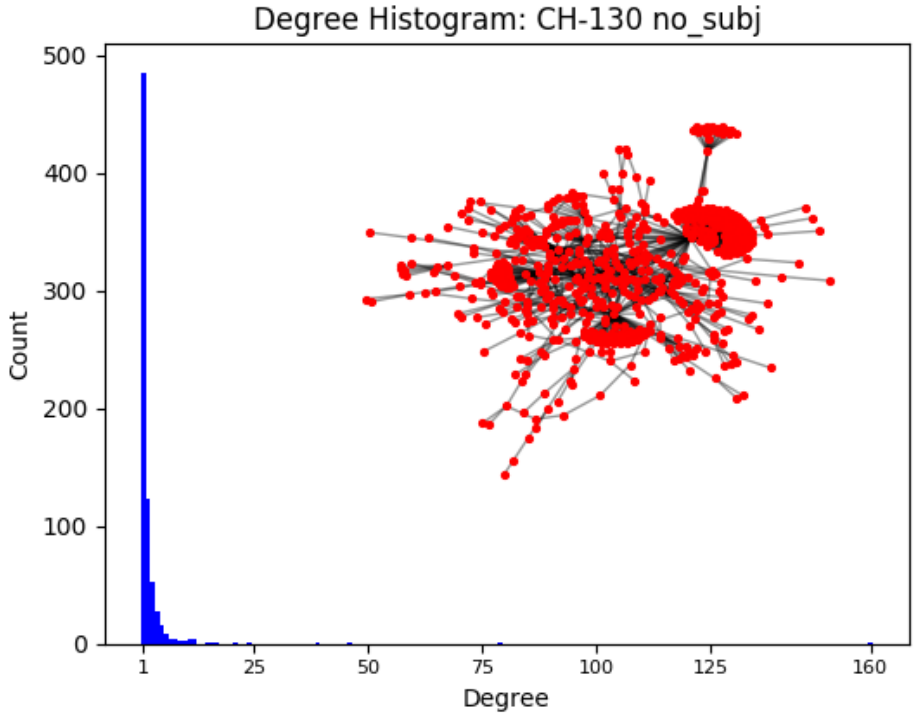


Figure 5.12.: CH no_subj graph and degree distribution

While such differences are somewhat visible to the eye, comparing graphs from only looking at them does not make for a very convincing argument, especially since graph visualization is computationally hard and relies on heuristic and iterative algorithms, so that the same graph can look very different depending on the algorithm and parameters used. Graph structure is therefore better captured and compared through connectivity measures based on significant subgraphs and the quantification of how easily the graph can be split into communities. In a graph of ten nodes, all ten can be connected with one another forming a complete graph (fig. 5.13), or none can be connected, forming an empty graph or a group of ten isolated communities (fig. 5.14). A complete and empty graph are in a sense structurally equal, because all nodes belong to either the same or each to their own community, meaning the structure cannot be changed or affected in any way by deleting any one node – there is no internal structure beyond being complete or empty.

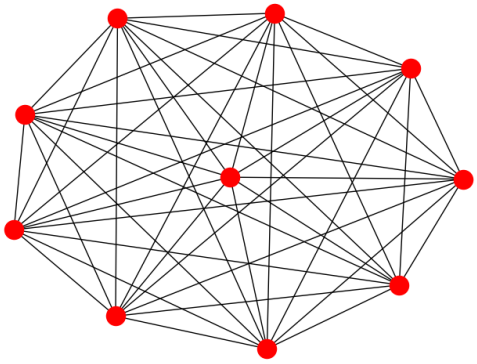


Figure 5.13.: Complete graph of 10 nodes

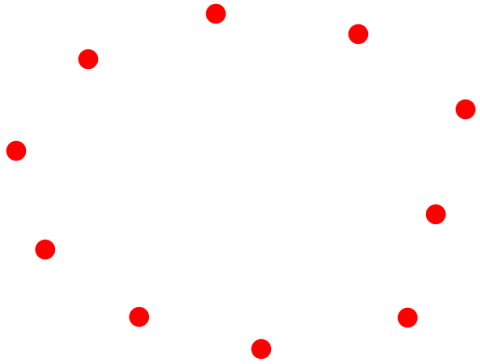


Figure 5.14.: Empty graph of 10 nodes

Incomplete, but not fully decomposed graphs, which are graphs with $0 <$ number of edges $< (\binom{n}{2} = \frac{n(n-1)}{2})$ where $n =$ number of nodes, can be more or less connected depending on the distribution (rather than the number) of edges and thus the tendency of nodes to cluster into communities: groups of nodes that are more tightly interconnected than others while at the same time having fewer connections to other communities. The term *community detection* refers to the NP-hard computation problem of detecting community structures as well as its solution through the use of a family of heuristic algorithms that approximate an optimal decomposition. This is relevant for the identification of crucial connecting points, rates of transmission between nodes, and the identification of building blocks and breaking points between them within complex systems. Algorithms of this kind are used in a variety of research fields in biology, social studies, and computer science. For an overview, see Fortunato (2010) and Schaub et al. (2017).

One metric of this kind that has been widely used is modularity, a measure in the range $[-1, 1]$ where 1 indicates highest modularity or partitionability, while lower values represent graphs that lack community structure (Fornito et al., 2016, 314).[23] One of the most widely used algorithms was developed by Blondel et al. (2008) and is called Louvain modularity[24] and works as follows (p. 4):

[23] Negative values represent graphs with fewer edges between communities than would be expected by chance. However, so far, I have not encountered a corpus-based graph with negative modularity. This may a matter of corpus size.

[24] The authors at that time all worked from UC Louvain, hence the reference. Google scholar citation count on 2019/08/04 was at $>$ 9300.

> "Our algorithm is divided in two phases that are repeated iteratively. Assume that we start with a weighted network of N nodes. First, we assign a different community to each node of the network. So, in this initial partition there are as many communities as there are nodes. Then, for each node i we consider the neighbours j of i and we evaluate the gain of modularity that would take place by removing i from its community and by placing it in the community of j. The node i is then placed in the community for which this gain is maximum (in case of a tie we use a breaking rule), but only if this gain is positive. If no positive gain is possible, i stays in its original community. This process is applied repeatedly and sequentially for all nodes until no further improvement can be achieved and the first phase is then complete. Let us insist on the fact that a node may be, and often is, considered several times. This first phase stops when a local maxima of the modularity is attained, i.e. when no individual move can improve the modularity."

The graphs in figs. 5.15 and 5.16 each have 10 nodes and 15 edges (the maximum number of edges in a graph of 10 nodes is $\binom{10}{2} = 45$), but show different structures and modularity values: Fig. 5.15 is a so-called Barbell graph, a symmetrical formation that consists of two equal-sized communities and a number of nodes connecting them, in this case two. As such, it is rather modular (Louvain modularity > 0.47), because assigning the two connecting nodes to one of the communities – which is similar to the deletion of an edge between the nodes and the community it is not assigned to and is how some other community detection algorithms work – suffices to split the graph in two. Fig. 5.16 shows two random graphs with modularity values of 0.1067 and 0.2711 respectively, where the first graph is already split into two isolated nodes and a larger community, which requires several operations in order to be split into further partitions, and the second can be partitioned by assigning two nodes to one of the groups, but the top right group is not fully connected as is the case in the Barbell graph, so the overall structure is less community-based or modular. In practice, small random graphs rarely seem to reach modularity values ≥ 0.3, while intentionally or systemically structured ones seem to reach higher values easily, as will be shown in the data analysis. Empty, complete, and path or circle graphs do not possess internal structure, because all nodes are connected to all other nodes in the same way (either all have exactly two neighbors out of n nodes, or zero, or n neighbors).

In conclusion, Louvain modularity provides a metric of the internal structuredness of a graph, does not naively reflect the number of nodes or edges in a graph,[25] and conveniently allows for a comparison between graphs within a single figure instead of a triangulation of various measures such as degree distribution, distances, number of communities etc. I will show in the next chapter that applying Louvain modularity to the Kobalt corpus based on the graph model developed in this chapter yields results that are in line with the hypotheses presented earlier. Of course, this alone is not an epistemologically safe argument for the validity of a method. A deeper linguistic validation clarifying whether coselectional constrainedness as a structural property of language(s) is well represented in modularity values is therefore necessary, but can unfortunately not be done within the scope of this thesis. Some qualitative remarks on that will be made in the discussion in chapter 7. One clear advantage of using a graph-metric like Louvain modularity is that it is strictly positivistic with respect to the data at hand and does not extrapolate to

[25] Which is not to say that it is not sensitive to these and will become relevant in the next chapter.

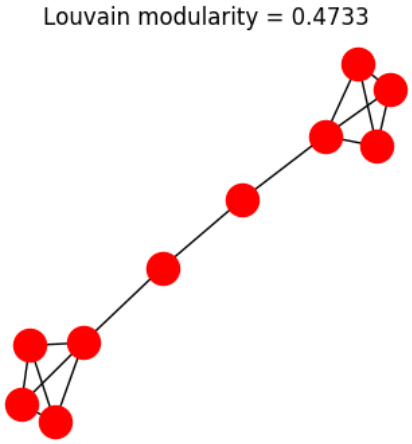


Figure 5.15.: Barbell graph of 10 nodes

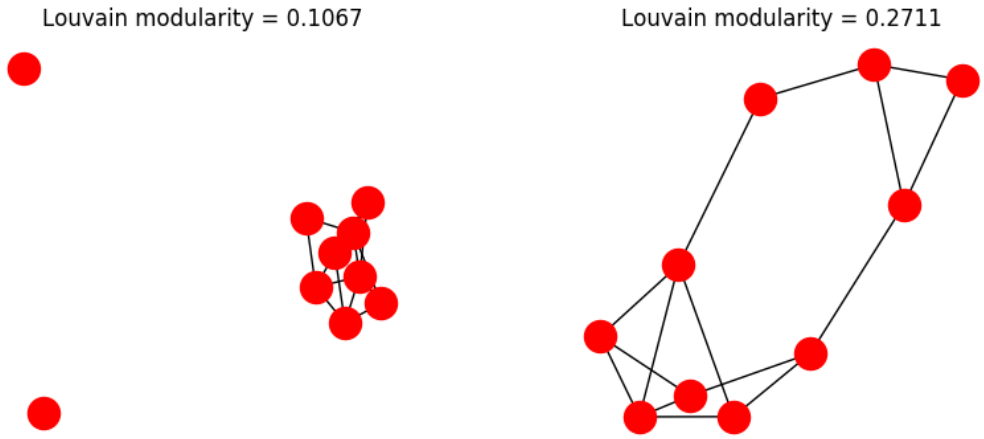


Figure 5.16.: Two random graphs of 10 nodes and 15 edges

an outer totality of unknown ontological status, and that it does not rely on unlinguistic assumptions of randomness or independence. How Louvain modularity compares to other measures of community or modularity cannot be assessed within the scope of this thesis, but is an interesting question for future research that requires a deeper understanding of the structural properties of a graph modeled from lexicosyntactic corpus information.

Louvain modularity is implemented in the Python NetworkX Community API (NetworkX: Hagberg et al. (2008), Community API: `https://python-louvain.readthedocs.io/en/latest/api.html`, developed by Thomas Aynaud), in neo4j/Cypher (Webber and Robinson (2018), `neo4j.org`), and in Gephi, a GUI program for network drawing and analysis (Bastian et al., 2009). Unless otherwise specified, the computations and visualizations in this thesis were done using Python with NetworkX, Community, and Matplotlib (Hunter, 2007) (this section); R (R Core Team, 2015) on RStudio (RStudio Team, 2015) with ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2018), reshape2 (Wickham, 2007), mgcv (Wood et al., 2016) and jsonlite (Ooms, 2014) (chapters 4; 6); and D3.js (Bostock et al., 2011) (chapters 5; 7).

5.4. Specified hypotheses

To specify the hypotheses derived earlier to the graph-based methodology, the following behavior is expected:

1. Louvain modularity is overall higher in L1 than L2, where L2 approximates L1-like values mmost in most advanced learners;

2. Learners show a u-shaped curve in Louvain modularity, where subcorpora including texts written by least and most advanced learners have higher modularity values than those in between;

3. Modularity is higher for more specific graphs (full graph < pp < vas_prep < vas_no_prep < no_subj);

4. Non-verb-specific graphs, full graph especially, show least of a trajectory from lower to higher onDaF scores of the included texts; trajectories are most defined and in line with the hypotheses (u-shape, modularity in L1 > L2) in vas_no_prep and no_subj; and trajectories are more similar for more similar graphs (full graph | pp and vas_prep | vas_no_prep and no_subj).

5.5. Summary

In this chapter, graphs as a knowlegde and information structure were introduced and used to build a model of lexico-syntactic coselection in Kobalt, where lexemes are represented by nodes and dependency by edges on five levels of specificity. A formal definition of the graph model can be found in appendix A.1. *Louvain modularity* was presented as a measure of graph modularity or connectivity that is not a trivial reflection of graph size, as measured in the number of nodes and edges, but represents a quantification of the interconnectivity and distribution of nodes and edges and the graph's decomposability into separate communities. It was argued that if every verb took exactly one argument, the graph would consist entirely of separate communities, and if every verb co-occurred with every argument, this would result in a complete bipartite graph, which would be

reflected in different modularity values. Higher constraints, and particularly distributional constraints, whereby classes of verbs interact more freely with some arguments than others, are represented by more tightly interconnected communities in the graph, translating to higher internal structure vs. an unstructured, random graph, again resulting in different modularity values. Modularity computation can therefore be mapped to a linguistic model of coselectional constraint and might work as a measure of the same. Finally, specified hypotheses for the behavior of the modularity metric in Kobalt were presented.

6. Results and validation of the graph-based analysis

At the beginning of chapter 3, the following research question was formulated: "(How) can the development of lexicosyntactic constraint be shown as a structural property in L2?". It was hypothesized that there is a development whereby learners reach higher constraint levels than they started at; that final levels are more L1-like; and that variance in L2 is higher than in L1. It was also hypothesized that the necessary process of diversification and specialization that occurs through the course of language acquisition is expressed in a temporary randomization of coselections, measurable in a u-shape or drop in coselectional constraint at intermediate stages of acquisition. A statistical analysis in chapter 4 has shown that randomization, diversification, and specialization are indeed visible, but that these processes also mask the development of coselectional constraint as it may exist for individual items. This is because, despite a generally high overlap in a core thematic lexicon, the number of identical coselections across and even within subcorpora is rather low.

It was shown that a statistical analysis is not ideally suited to capture the expected changes, since it can only compare factor combinations (identical lexemes/coselections or categories of lexemes/coselections). It was further argued that, even if there had been more identical coselections, results from a statistical analysis would be still be difficult to interpret. Without a quantiative model of the *idiom principle* (Sinclair, 1991) the interpretation is lost in a huge combinatorial space that is inherent in the potential coselection of unique verb and argument lexemes even in a smallish corpus like Kobalt. Against this background, randomness does not serve as a plausible baseline to decide whether the set of realized coselections should be considered constrained or not.

As an alternative, a graph-based model was introduced in chapter 5. Graphs differ from factor combinations in at least three relevant ways:

- They model more information through the inclusion of relational information between all items in the model;
- They abstract from the identity of elements and thus allow for a comparison of diverse items in the same relational space;
- Metrics that measure aspects of the graph are, like descriptive statistics, strictly positivistic and do not infer to outer populations, totalities, or presumed probabilities.

It was suggested that a graph model might be more representative of the expected effects, and that a community clustering algorithm and an analysis that provides a unified measure, *Louvain modularity* (Blondel et al., 2008), could be capable of showing them more clearly. It was also suggested that this would be preferable given that a triangulation of a number of statistics leaves a high degree of uncertainty, unless it is precisely validated for all interacting effects. This in itself is a complicated task would require abundant data that is not available.

The aim of this chapter now is two-fold: It still sets out to answer the research question of whether a restructuring of coselectional constraint can be measured in learner text, now through a graph-based data model.

At the same time, since graph-based metrics (unlike graphs for visualization) are virtually unused in present day linguistics, the method itself requires validation. Typical confounding factors like individual effects vs. group and grouping effects, corpus size, and text length need to be controlled for in order to gain a better understanding of the measure and its mechanics in corpus data. This is both a contribution to the methodology of the development of methods in corpus linguistics, where methodological validation is currently less discussed; and an assessment of the measure itself and its utility in application to small to medium-sized corpora.

Since both aims of the chapter are marked by theoretical and empirical uncertainty, this thesis can only serve as a first approximation. This is why this chapter reports results for Louvain modularity values for coselections of graphs that include *all* VAS with object, subject, and predicate type arguments (OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, PRED, SUBJ, SUBJC) as defined in the graph specificities in chapter 5, rather than looking into the coselectional constraint of individual slots. This conflates phrasal with nominal arguments and could be criticized as a lack of linguistic differentiation. At the same time, very little is known about the details of coselectional preferences or constraints in general, and there is nothing that indicates that phrasal arguments do not underly coselectional constraints and the same development.[1]

It is likely that the inclusion of verbal arguments renders graphs overconnected (understructured). This would yield weaker results and needs to be examined in future research, along with analyses of specified VAS or VAS slots, which may also require a more fine-grained semantic differentiation of slots (such as the unergative-unaccusative distinction) or verbs (such as complex vs. simplex verbs).

Similarly, pronouns are included in this analysis. Considering more semantically oriented approaches to argument selection constraints like Plank (1984), and a statistical perspective, pronouns should be treated with caution. They are deictic and thus semantically flexible, and they are frequent. Thus they are – in a semantic approach – unlikely to exhibit specific constraints, while also skewing results for the less frequent lexemes. At the same time, from a phraseological or form-oriented perspective, as well as phonotactically, and certainly in real world distributions (*She gave birth to a baby boy*), it is not implausbile to assume that pronouns *do* have coselectional preferences, or that verbs prefer some pronouns over others. In addition, this raises the question of whether all pronouns should be excluded. With the case of indefinite pronouns like *jede/r* ('every') or *alle* ('all'), this would exclude a number of frequent subject lexemes in learners that are not frequent in L1 (see section 6.2), thus deleting interesting information from the model. Within-category distincitons, such as the in- or exclusion of certain lexemes by semantic criteria, were generally avoided. The model that has been discussed in chapters 3.2.2 and 5.2 instead relies on structural filtering based on the annotations reported in chapter 3.2.

In all respects, where skewing of the data had to be accepted for a first application of the measure – whether due to the lack of an existing linguistic differentiation and theory

[1] In fact, most of collostructional analysis and other statistical approaches is applied both to nominal and to verbal or phrasal slots (see section 2.1.2); and statistical approaches to collocation extraction are often positional (n-gram-based) and thus conflate even more diverse linguistic categories. This goes to show that the assumption is that coselectional preferences may exist across category boundaries is widespread and common.

of coselectional constraint or due to processing constraints – it was attempted to keep results on the conservative side. Thus, the most specific graphs are *at least* as structured or modular as presented, and would gain in modularity through further differentiation. In other words, results are likely *less* in line with the hypotheses than they could be if further differentiation was applied.

It will be shown that these limitations are of a more theoretical concern, though, and that modularity values shift with graph specificity for each onDaF and language group, but do not differ in distribution or trajectory by specificities except for the most specific graphs. This will be discussed in section 6.2. While trajectories slightly differ *within* groups, there is nothing in the data that indicates that a different level of specificity (aside from a complete separation of slots) yields different patterns *between* groups.

Since Louvain modularity is a heuristic measure that depends on the order in which neighboring nodes are fed into the maximizing function, I have computed modularity values for each subgraph 350 times and used the maximum of those, effectively giving a lower bound for modularity values (each graph is *at least* as structured as the corresponding modularity value suggests).[2].

Louvain modularity can be computed for weighted and unweighted graphs. Graphs in this model are weighted, where edge weight signifies frequency of co-occurrence. However, in this data, weighted and unweighted modularity differ mainly in absolute values, where weighted modularity is higher for the same sample. Trajectories are basically identical. This may be due to the Zipf-distribution, where with the large number of hapaxes, most of the graph is identical for weighted and unweighted edges, and weight concentrates on few items which are also tightly interconnected (consider for example coselections of the verb *haben* 'to have', of which there are many, but which are also recurrent). In the spirit of keeping results on the conservative side, only unweighted modularity will be reported in this chapter. A plot comparing weighted and unweighted modularity is included in the appendix (A.3).

Results are first reported for splits of the data by onDaF-ranges as they were defined earlier (section 6.1). For validation, a comparison with smaller onDaF ranges is included. Unfortunately, the distribution of texts across onDaF ranges makes it impossible to isolate grouping from corpus size effects, and results are inconclusive. They are still reported in section 6.3.1.2, because they corroborate a point that will come up again in section 6.3.2.2, namely that while analyses of individual texts (6.3.1.1) and corpus sizes ≥ 9 texts yield consistent results in this dataset, corpus sizes of 5 and 6 texts do not. This will be discussed in section 6.3.3.

Furthermore, three sampling techniques that are currently uncommon in corpus linguistics are introduced and discussed: An out-of-sampling for an assessment of group vs. individual effects (section 6.3.2.1); a sliding-window-sampling of 5-, 10-, 15-, and 20-text-

[2]The appendix includes an overview of approximations of the limit in 500 iterations (A.2). It appears that the limit is reached more or less steadily after around 100-150 iterations in this type of text and corpus size, and that clear patterns by group emerge even sooner. This is due to the fact that not all nodes in a lexicosyntactic graph are different in a graph-structural sense: Some are part of similarly structured communities, and many are connected to the graph through only a single edge (hapaxes). In that case, choosing either one out of a similar set as a starting point may lead to the same division of communities (modularization). The larger the graph, however, the more likely it is that those nodes that lead to the optimal modularization are missed by chance in fewer iterations. Since minor changes can still occur until after 300 iterations, 350 iterations were chosen as a limit. While the exact values may still not have been reached in this, practical constraints do apply: With 350 iterations, computations on a 24x2.7GHz server at full capacity took about 8 days in total for all data splits.

windows for a simulation of a more balanced dataset, to gain an understanding of corpus size effects, and for an estimate of the continuity of an implied trajectory (section 6.3.2.2) and a sampling of a fixed number of verb-argument structures for a normalization of text length (section 6.3.4).

Results overall show major agreement with several hypotheses:

- Graph specificity determines graph modularity, and trajectories are more similar for the verb-specific graphs (except for no_subj, see below);
- L1 graphs are more modular than L2 graphs;
- L2 graphs are more modular at advanced vs. earlier stages;
- A u-shaped development exists in BEL.

Hypotheses were not confirmed in two ways:

- No_subj graphs are not simply more modular than other verb-specific graphs, but showed distinct trajectories, and are more modular in some learners vs. L1. This will be discussed in section 6.2.
- A u-shaped development cannot be observed in CH in most analyses. This will be discussed in chapter 7.

6.1. Results by onDaF-group

In what follows, Louvain modularity has been computed separately for each subgraph based on graph type and grouping, and results will be presented viewing the onDaF scores as a progression or time series.

As was discussed earlier, I divided Kobalt into subcorpora by language and onDaF groups, where 75, 95, 115, 130 and 160 refer to the group of texts whose authors scored < 75, 75-94, 95-114, 115-129, and 130-160 points. Ranges were chosen pragmatically around the original corpus data (see chapter 3.2.1 for a detailed introduction of the data). These happen to vaguely correspond to CEFR-related cut-off points of the English equivalent (Eckes, 2017),[3] but theoretical claims about precise CEFR-localization are not intended. Rather, data will be referred to as *lower-intermediate* (75, 95), *higher-intermediate* (115, 130), and *advanced* (160).[4].

Since the distribution of texts is unbalanced across onDaF scores in Kobalt[5] – there are 24 texts in CH-95 and only 10 in CH-160, 27 in BEL-95 and 11 in BEL-160 – and

[3] `http://www.fremdsprachenzentrum-bremen.de/fileadmin/autor/datein/Symposion_2017/Praesentationen/AG4_Eckes_Symposion2017.pdf`

[4] Obviously, it is not intended to say that a learner switches from 'higher-intermediate' to 'advanced' based on a single onDaF point. Dividing a continuous scale into discrete classes always carries the problem of implying a qualitative jump through minor quantiative changes. While such dynamics may exist, they are not implied here. Rather, since hypotheses were made with learners at early, intermediate, and advanced stages of target language acquisition, some categorization is required.

[5] This is due to the original data collection of the Kobalt project, which set out to build a small, but balanced and deeply annotated corpus of essays written by learners within a narrow range of onDaF points at roughly B2-level (115-130 points, see Zinsmeister et al. (2012)). The 111 texts outside of this score range, which are included in the analysis in this study alongside the core Kobalt data, were collected incidentally from what was available and without further balancing. The Kobalt project set out to collect 20 texts in the onDaF score range of 115-130 per language, which is labelled as the onDaF

since Louvain modularity is expected to interact with corpus size, results are reported for 10-text-samples from each subcorpus. A comparison of five such samples per subcorpus is shown in fig. 6.1, where all samples partially or fully overlap since there are fewer than 50 texts in each subcorpus. For better legibility, fig. 6.3 shows only the modularity values for the most specific graph types.

Results relate to the hypotheses in the following ways:

- Confirmed: Graph specificities determine absolute values of modularity. Higher specificity leads to higher modularity;
- Confirmed: Verb-specific graphs have more distinct patterns compared to full graphs;
- Inconclusive: Variance is higher in L2 vs. L1 in some graphs, but not all (see also fig. 6.2). This may be an artifact from the larger diversity of the sample (20 texts in L1-subcorpus vs. 11 in CH-160, 10 in BEL-160, etc.);
- Confirmed: Modularity is higher in L1 and L2 for all specificities (except no_subj, where CH-160 has higher modularity than L1 and two more L2 corpora have modularity values similar to L1);
- Confirmed: Modularity is higher in more advanced learners (BEL-130; CH-130, CH-160) compared to early intermediate ones for most graphs (except BEL-160, and CH-160 (no_subj) where modularity drops;
- Confirmed: BEL-learners show a u-shaped trajectory in all verb-specific graphs; At odds: CH-learners do not.

Three aspects require further investigation: The aberrant behavior of the no_subj graphs, which will be discussed in section 6.2; the absence of a u-shaped development in CH learners, which will be discussed in chapter 7; and the drop in modularity in BEL and in some CH graphs in the most advanced learners. This can partially be explained through text length/text structure as will be shown in section 6.3.4.

For an acceptance of the major hypotheses, a careful validation against two sensitive aspects is still required: Corpus size, which here is controlled in terms of the number of texts, but not tokens; and grouping, where it needs to be shown that effects between onDaF groups are indeed larger than inter-individual differences across onDaF ranges,[6], and that a grouped analysis is indeed both superior to an analysis of individual texts and valid with respect to the group size and onDaF range chosen.

130 group here. Three authors of CH-texts included in the original corpus only reach 114 points and were reassigned to onDaF 115 in this analysis, hence there are only 17 texts in CH-130. Despite this limitation, the corpus is linguistically rich and highly valuable, since it is more controlled in terms of topic (or rather, prompt), writing conditions, and learner cohort than other available corpora, certainly cross-sectional ones of German SLA. In addition, the smallish size allows for more and deeper manual and semiautomatic annotation. Since unbalanced, sparse, and irreplicable data are a recurrent theme in corpus linguistics, I will discuss in this chapter ways to validate results quantitatively from within the data through sampling and groupwise comparison. This does obviously not create more or more balanced data, but it does provide deeper insight.

[6] In other words, an estimate of the influence of inter- and intra-individual variance is required to define a lower bound for a corpus size at which the signal of an emergent phenomenon like coselectional constraint becomes strong enough against the noise of individual variance.

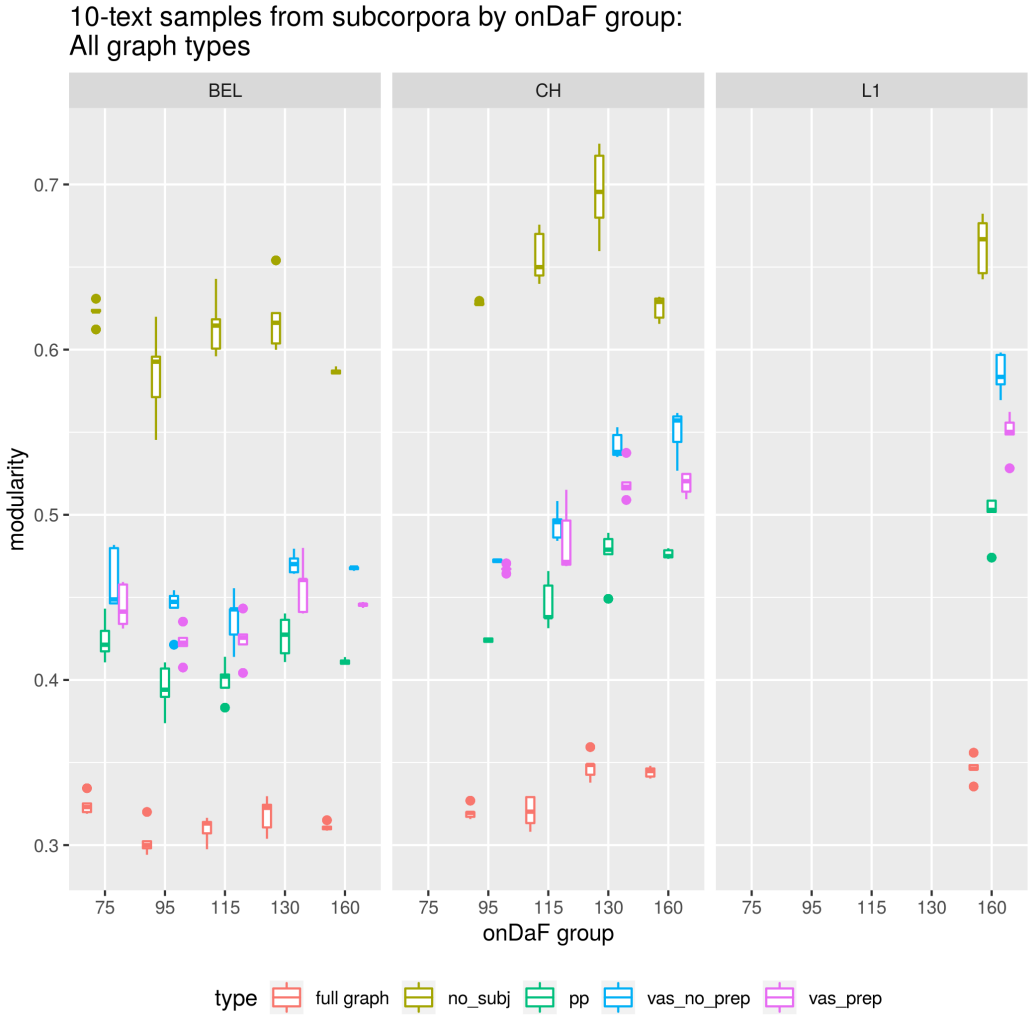


Figure 6.1.: Modularity in 10-text-samples from subcorpora by language, 5 samples per subcorpus. Modularity is higher in advanced learners than early- and high-intermediate ones in CH, and in BEL-130 (but not BEL-160) in all specificites except no_subj. Modularity is higher in L1 compared to L2, except for no_subj. A u-shaped development exists in BEL, for most specificities, but not in CH. CH modularity fits almost neatly between BEL and L1.

Unfortunately, effects from grouping and corpus size cannot be isolated in Kobalt, because the data cannot be split in such a way that both corpus size and onDaF score ranges are balanced across the dataset while keeping sufficiently large samples at the same time. There is a trade-off between linguistically based grouping, corpus or sample size, and balance: Samples can be within a small onDaF range but are then restricted to 6 texts per subcorpus (see the onDaF10 analysis in section 6.3.1.2); or they can be any chosen size and equal in size even at the lower or higher end of the onDaF range, but fluctuating in variance of onDaF ranges in a sliding-window analysis (see section 6.3.2.2); or they can be large enough and controlled in terms of onDaF criteria, but with smaller subcorpora at either end of the onDaF range, as in the original grouping presented in chapter 3 and 4. Validating for grouping effects is not primarily relevant for a confirmation of the results

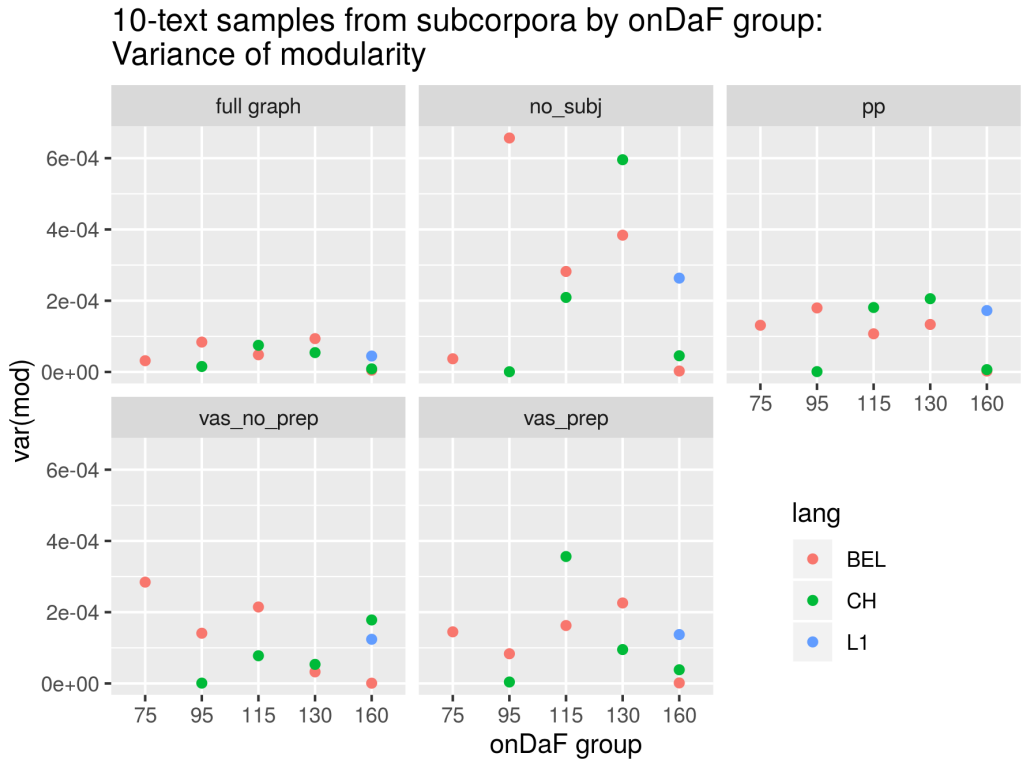


Figure 6.2.: Variance of modularity in subcorpora by onDaF group. Variance is higher in L2 vs. L1 in some graphs, but not in all.

from this grouping in larger samples, but more so to verify that it does provide comparable results representative of a learning trajectory. This is important for any more qualitative study of coselections in the corpus that may follow later and that would be unable to consider all 67 sliding windows in BEL, but can handle five groups in an onDaF grouping.

In addition, the variance in text length particularly in the BEL corpus is quite large, since text length grows linearly with onDaF scores in BEL-learners. This raises doubts regarding the comparability between texts written by early intermediate and very advanced learners.

Consequently, an internal validation of results will be attempted through analyses of systematic splits of the data based on corpus size, grouping, and text length normalization in the remainder of this chapter.[7] I will refrain from further predictions for the individual splits, walking through them in a more exploratory fashion. This is, on the one hand, to avoid circularity and what might be likened to p-hacking in statistics, i.e. excessive testing that has a high chance of producing random chance results congruent with the initial hypotheses. On the other hand, further hypothesizing is redundant in this study, since the method was developed on previously seen data. Any further testing would thus be prone confirmation bias.

Including a careful validation in the presentation of a new method is further relevant to the systematization of methodological development in corpus linguistics. There are cur-

[7]For a discussion of the benefits of internal validation, see Steyerberg and Harrell (2016); Steyerberg (2018).

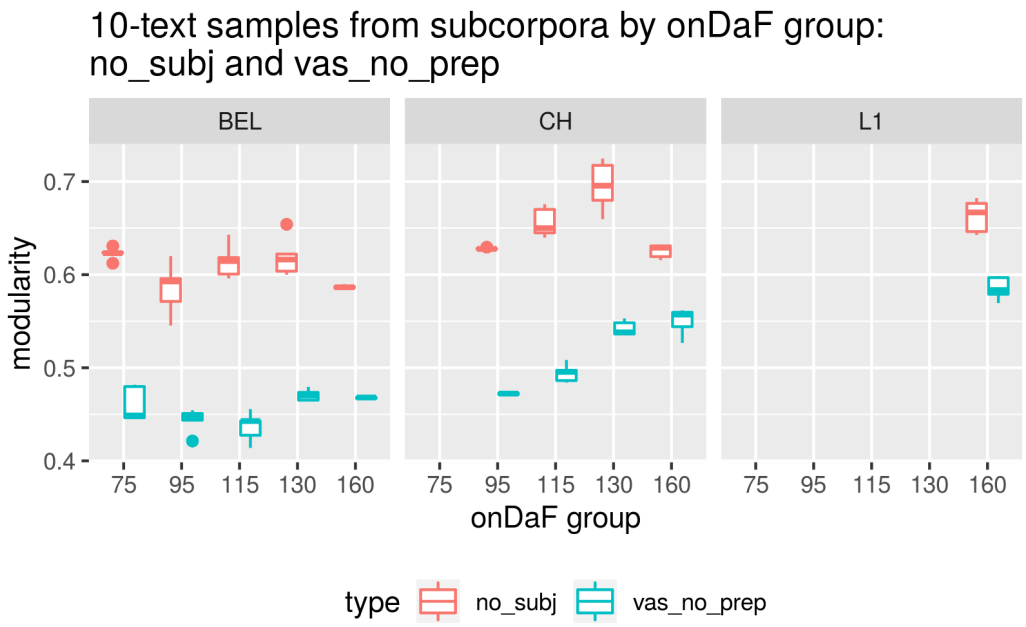


Figure 6.3.: Modularity in 10-text samples from onDaF-based subcorpora by language, 5 samples per subcorpus, no_subj and vas_no_prep only. A u-shaped development is clearly visible in BEL, but at different onDaF groups for the two specificites. CH shows no u-shaped development. Modularity is higher in L1 for vas_no_prep, but not for no_subj vs. CH-130 and CH-95.

rently no best practice guidelines for sampling or other methods for internal validation in corpus linguistics in general,[8] much less for unbalanced and sparse corpus data. Of course there is a multitude of sophisticated sampling methods being discussed in mathematics, engineering, and the natural and social sciences. But those cannot be applied to text without consideration of the common caveats of quantitative linguistics as they are found in statistical testing, too: The Zipf-distribution of lexical items and its influence with potentially non-converging frequency limits (i.e. potential failure of the central limit theory, concept of probability) in interaction with text length, corpus size; high levels of inter- and intra-individual variance; and text-linguistic factors, such as the holistics of a text, the failure of randomness assumptions concerning the order of appearance (first half of a text vs. second half, conditional probability from long distance dependencies), and so on.

This chapter is therefore also meant as a first attempt to compare and systematize different sampling approaches for data of this kind. Sparse, unbalanced, and incomplete data are a reality in many subfields of linguistics, most notably in historical or less documented languages. Therefore, finding agreeable approaches towards internal quantiative validation is desirable for methodological clarity and for the extension of the scope of research questions towards such data. Even subfields where data is potentially abundant such as

[8] There is some work on sampling strategies in corpus compilation, such as Evert's library metaphor (Evert, 2006) and the discussion of represetativeness of a corpus (Biber, 1993). But this is a different kind of sampling, if one wants to call it this, namely a sampling in the choice of data (what to include) from a population, not a validation of existing data to account for as much of the variance as possible without external modification (how do observations change in different splits of *the same* data).

SLA research, data collection, particularly of controlled data such as is used here, remains a time consuming and resource intensive task, and approaches towards making sparse data more usable are relevant to grow sustainability and research efficiency.

6.2. Graph specificity and subjects in L1 vs. L2

To shortly recapitulate the model of graph specificity presented in chapter 5.2:

1. Full graphs contain all lexemes and dependencies of a text or subcorpus;
2. *pp* and *vas_prep* graphs contain all verb argument structures including prepositions and their complement head nouns, inclusive of all PPs attaching to the verb (pp) vs. only OBJP-labeled PPs (vas_prep). OBJP are prepositional objects that are lexically specified in the verb signature, generously erring on the side of including potential OBJP (see sections 3.2.2 and 5.2 for more details);
3. *vas_no_prep* contains only VAS and their arguments, where noun complements to OBJP prepositions are included, but not the prepositions themselves. This is to avoid transitivity between verbs and argument lexemes where two verbs connect to the same preposition and through that to all preposition complements, whether attached to the specific verb or not, which creates hyperconnectivity.
4. *no_subj* is the same as *vas_no_prep*, but exclusive of all subjects. Subjects here refer to semantic subjects in active/passive distinction: Subjects in active verb constructions are left out, while passivized objects ('*the decision* has been made') are modeled as OBJA. Distinctions based on verb semantics (unaccusative/unergative) are not made. See chapter 5.2 for discussion.

The boxplot of modularity values in 10-text-samples from onDaF groups in fig. 6.1 showed that higher specificity correlates with higher modularity, as was predicted. As was also predicted, trajectories over the implied timeline vary in correlation with graph specificities. Most notably, a clear distinction was observed between full graphs and verb-specific graphs, where trajectories were more clearly defined in the verb-specific graphs; and in a shift between vas_no_prep and no_subj, that had not been predicted. L1 modularity values were higher for verb-specific graphs, but not for full graphs. This suggests that full graphs, with their inclusion of function words, are hyperconnected relative to the model of coselectional constraint.

Fig. 6.4 shows the distribution of data points from 10-text-samples for the three verb-argument-specific graphs. Interestingly, however, the no_subj graph differs not only in absolute modularity, but also in trajectory and variance: Fig. 6.2 had also shown that variance is more than twice as high in no_subj compared the other graphs in six out of ten subcorpora.

It appears thus that the difference between the no_subj and the other graph types is of a qualitative kind. The argument in favor of distinguishing between a graph inclusive of subjects and one that has only non-subject type arguments was that most of grammar theory, as well as semantic categorization, suggests that subjects are unlike other argument types: They differ in the syntactic relations they enter (agreement/mutual c-command vs. government), they do not form semantically complex meanings to the same degree (*to play video games* vs. *to play music*) and thus underly fewer coselectional constraints in the argument of (Plank, 1984). This would suggest that, since they appear to be semantically

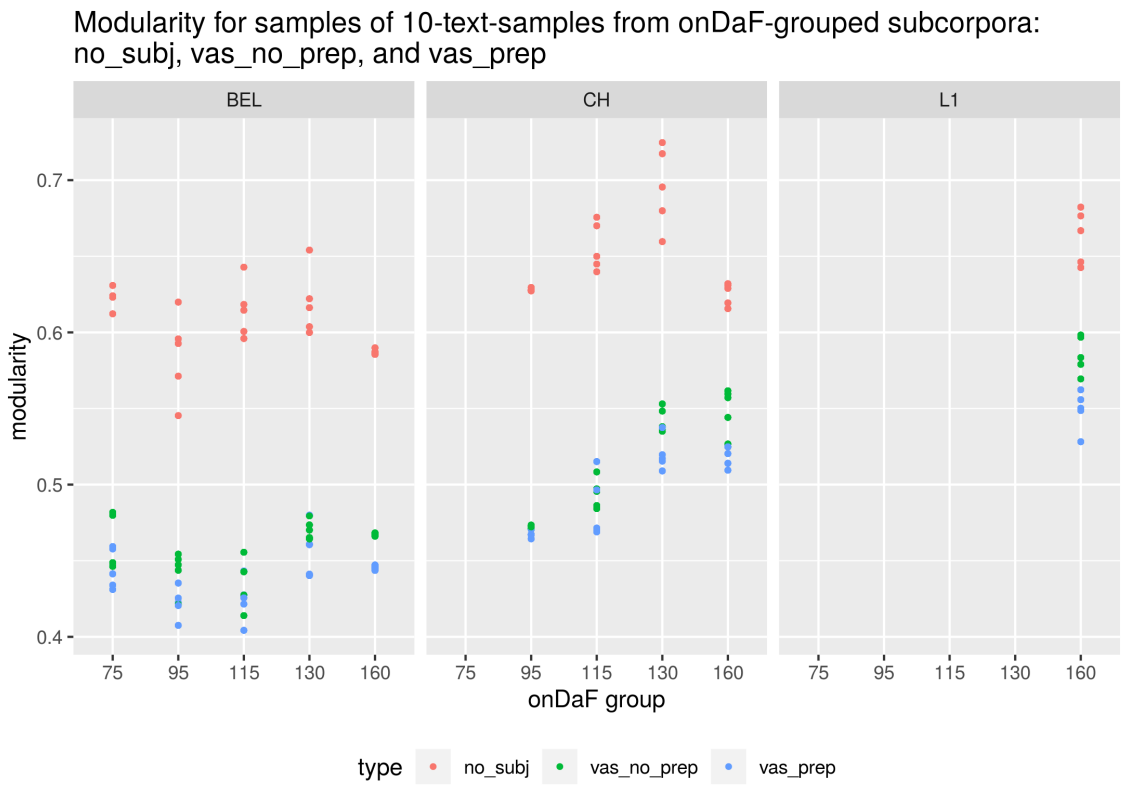


Figure 6.4.: Modularity in 10-text samples from onDaF-based subcorpora by language, no_subj, vas_no_prep and vas_prep, 5 samples per subcorpus. While distributions from vas_no_prep and vas_prep are rather similar and partially overlap, no_subj shows much higher modularity, greater variance, and a larger drop in modularity in the most advanced learners. L1 modularity is on par with CH- and some BEL-samples in no_subj, but only overlaps with two samples in CH in vas_no_prep (CH-130, CH-160 and one in vas_prep (CH-130).

less specific, learners have an easier time using them. However, results in chapter 4 suggest differently: Learners were shown to use more unique OBJA lexemes than SUBJ results, and many fewer SUBJ lexemes than native speakers in total. Fig. 6.5 shows this pattern in individual texts. In the L1-group, variance is larger and overall more unique lexemes are used as subjects vs. accusative objects (consistens with Plank's analysis), while for the learners, the opposite is true. In BEL-160, where texts are 1.5-2 times the length of the average L1 text, the median number of unique object lexemes is significantly higher in a two-sided t-test ($p<0.02$), while the number of unique subject lexemes is not.[9]

Interestingly, more half of the learners (above the median in the plot) use more unique lexemes in both SUBJ and OBJA slots than half of the native speakers (below the median line in the plot), and one BEL-115 learner almost three times as many OBJA lexemes as the L1 average. If these were mostly hapaxes, they should increase modularity. But since modularity is also lower in these groups compared to the edges of the onDaF range,

[9]It is significant at $p<0.05$ in a one-sided t-test based on H_1 = unique OBJA lexemes in BEL > unique SUBJ lexemes in L1, but barely so ($p=0.0498$). Significance is a problematic concept in lexis, but since no comparison of Zipf-distributed lexemes is undertaken, but of two categories that were expected to be similarly distributed, I believe an application of the t-test is not misleading here.

those lexemes must be woven into the graph in a way that creates higher connectivity. This suggests that new lexemes are preferably used with pre-used verbs, rather than introducing verb-argument-complexes, where the latter would raise specialization.

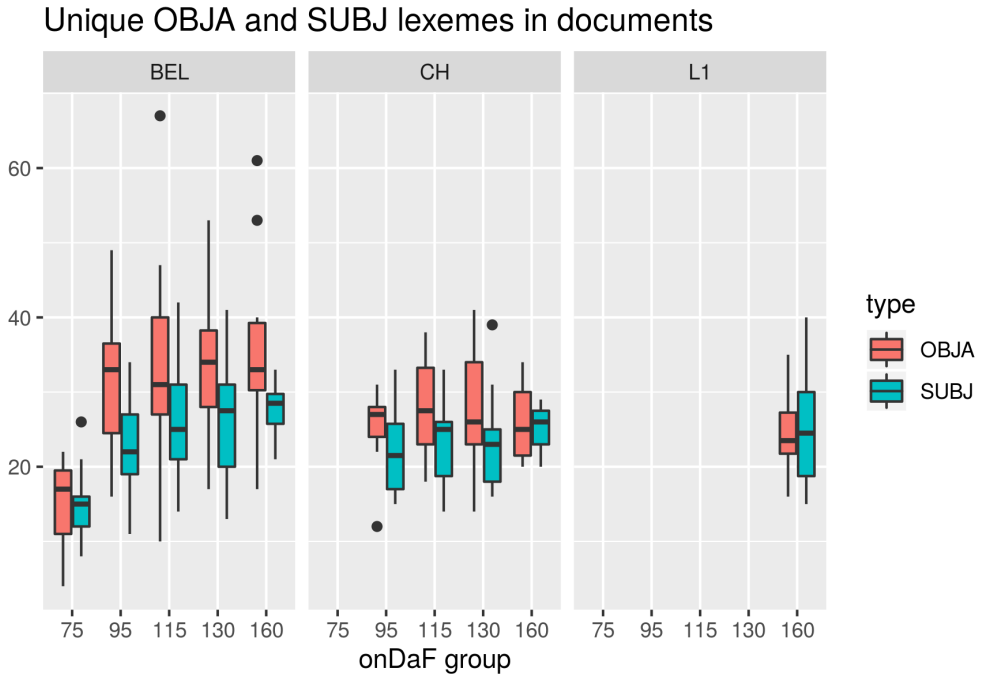


Figure 6.5.: Number of unique subject and object lexemes in documents. Learners use more unique lexemes in both slots than many native speakers. The boxes for SUBJ and OBJA differ in L1 and L2: native speakers use slightly fewer, but mostly a smaller range of OBJA compared to SUBJ lexemes, while learners consistently use more, and a wider range of OBJA lexemes than SUBJ lexemes.

In all three language groups, the number of unique SUBJ lexemes grows within a linear band with each unique OBJA lexeme, more steeply so in the BEL and L1 groups than CH (see fig. 6.6. However, as fig. 6.7 shows, a clear difference exists in interaction with text length: Learners introduce new OBJA lexemes more frequently than SUBJ lexemes, while native speakers introduce SUBJ lexemes more frequently. This translates well to the concept of slot-specific coselectional constraints. At the same time, it is important to see that large variance in L1 exists, such that L1 and L2 ratios overlap for some texts of similar text length.

A lexical analysis further corroborates the interpretation that there are meaningful differences in the use of SUBJ and OBJA in learners vs. native speakers. Out of the 25 most frequent SUBJ lexemes in the L1 corpus, ten are not used frequently in some or all L2 subcorpora (fig. 6.8).

- Three of those lexemes are relative or demonstrative pronouns, namely *welch* ('which'), *wer* ('who') and *dies* ('this'), which point towards syntactic differences between learners and native speakers;
- The other ones seem to express text structural L1 preferences, with four of them belonging to a terminology of argumentation (*Frage* ('question'), *Problem* ('problem'), *Vorteil* ('advantage'), *Begriff* ('term, concept'), see fig. 6.8.

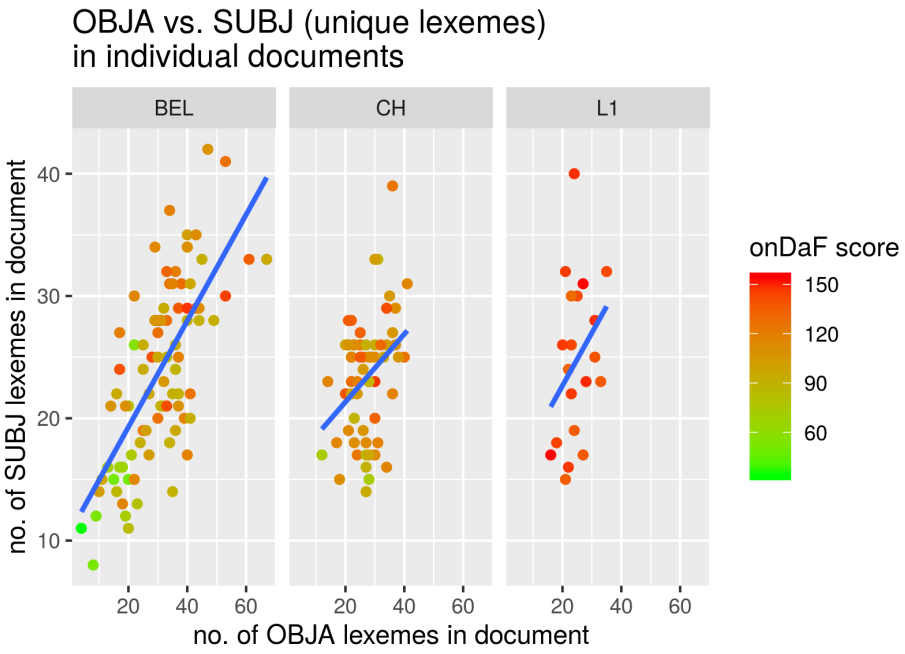


Figure 6.6.: Unique OBJA and SUBJ lexemes in individual documents. More advanced learners (higher onDaF score – orange and red points) are grouped above the regression line in both learner groups, suggesting a more L1-like ratio.

Fig. 6.9 shows lexemes that are frequent in CH and BEL, but not in L1. Two groups can be distinguished here:[10]

- Indefinite pronouns: *jede* ('each, every'), *alle* ('all'), *viel* ('many')
- Generic nouns: *Leute* ('people'), *Zeit* ('time'), *Situation* ('situation'), *Welt* ('world')

Both of these groups suggest low specificity through generalized statements. Lexemes that are frequent in only one of the L2 groups are, with only a few exceptions, among the most frequent only in one subcorpus of the language. Despite this, they appear to synthesize into a coherent picture of topic differences:

- Frequent in CH, but not BEL or L1 are a number of nouns related to social actors and societal developments *Arbeiter* ('workers'), *Regierung* ('government'), *Wohlstand* 'prosperity', *Chance* ('chance, opportunity'), *Umwelt* ('environment'), *Technik* ('technology'), *Unterschied* ('difference'), *Bedingung* ('condition'), *China* ('China'), and *Gesellschaft* 'society', which is among the 25 most frequent in all CH-subcorpora;
- Also frequent in CH, but not BEL or L1 are the indefinite pronouns *manche* ('some') and *andere* ('other');
- Frequent in BEL, but not CH or L1 are *Junge* ('boy, (the) young'), *Oma* ('granny'), *Student* ('student'), *Großeltern* ('grandparents'), *Freiheit* ('freedom'), *Arbeit* ('work'), *Technologie* ('technology'), *Krieg* ('war'), *Gedanke* ('thought'), *Frau* ('woman');

[10] While all of these are used in L1, they appear more often in other slots (most interestingly 45 times as PN, argument to PP, and 72 times as DET, 37 times as ADV, and 18 times as OBJA/OBJD) and still 30 times as SUBJ, which is not among the most frequent 25 lexemes in SUBJ.

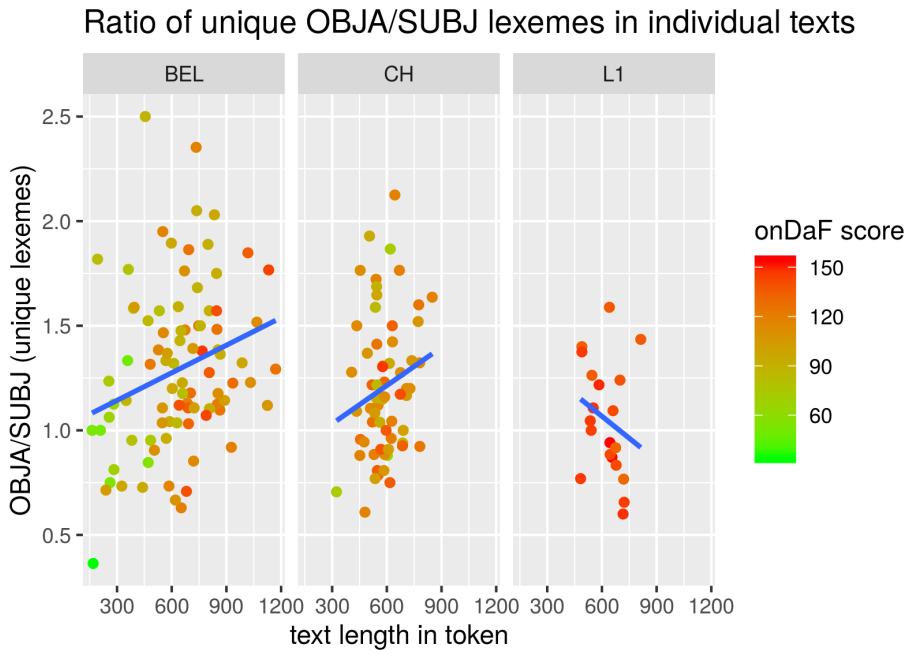


Figure 6.7.: Ratio of unique OBJA/unique SUBJ lexemes in individual documents by text length. onDaF-scores are now less predictive. In L1, new SUBJ lexemes are introduced more frequently than new OBJA lexemes, suggesting a lower upper bound for OBJA (higher coselectional constraint). In learners, OBJA lexemes are introduced more frequently than SUBJ lexemes.

- Lexemes that are frequent in BEL and L1, but not CH are *Mensch* ('human, man'), *Leben* ('life'), *Kind* 'child')

This suggests that CH-learners discuss more abstract or societal topics (what has changed from previous generations *in society*), while the BEL-learners take a more personal perspective (what has changed for my family/people like myself from previous generations?).

The 15 lexemes that are among the 25 most frequent in at least one subcorpus of each language group are

- pronouns and determiners: *d (der, die, das)* (demonstrative or relative pronoun), *es* ('it' as pronoun and expletive), *Sie/sie* ('they, she, (honorific) you'), *wir* ('we'), *ich* ('I'), *man* ('one'), *was* ('which, what');
- nouns used or closely related to the prompt: *Generation* ('generation'), *Jugend* ('youth'), *Jugendliche* ('youth, young people'), *Kind* ('child'), *Eltern* ('parents');
- *Leben* ('life'), *Mensch* ('human, man'), and *Problem* ('problem, issue').

11 are among the 25 most frequent in all subcorpora (*Kind* ('child'), *Leben* ('life'), *Mensch* ('human'), *Problem* ('problem') are not). This leaves a list of grammatical subjects (expletive *es*; demonstrative, relative, personal, and indefinite pronouns) and prompt-driven lexical nouns.

With this, while learners and native speakers use subjects functionally and in relation to the prompt, where they diverge, they do so somewhat systematically:

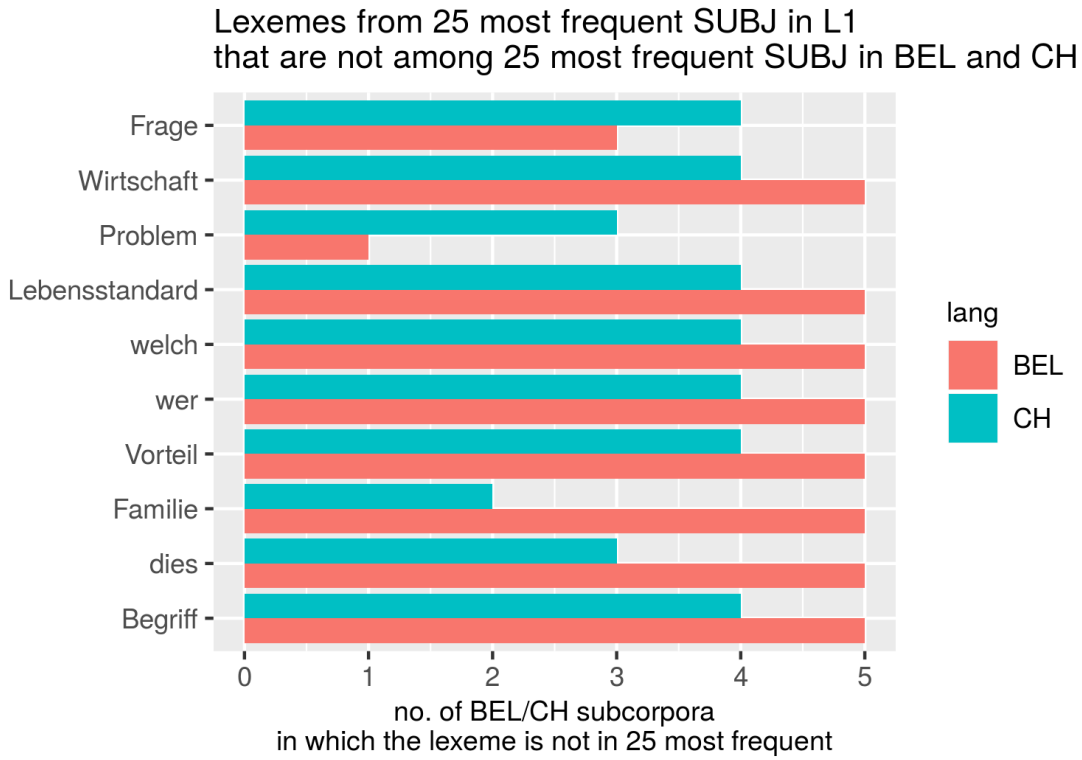


Figure 6.8.: Frequent subjects in L1 but not L2: *Frage* ('question, issue'), *Wirtschaft* ('economy'), *Problem* 'problem, issue', *Lebenstandard* ('living standard'), *welch* ('which'), *wer* ('who'), *Vorteil* ('upside, advantage'), *Familie* ('family'), *dies* ('this, that'), *Begriff* ('notion'). For example *Frage* ('question, issue') is among the 25 most frequent SUBJ lexemes in L1, but not in any of the 4 CH-subcorpora, and not in 3 out of 5 BEL-subcorpora. *Wirtschaft* is among the 25 most frequent SUBJ lexemes in L1, but not among the 25 most frequent subjects in any of the L2 subcorpora.

- Native speakers use SUBJ lexemes that establish an argumentative, functional frame (*Frage, Problem, Vorteil*);
- Learners seem to frequently use subjects in a functionally different, perhaps more semantically driven way:
  - to express who does something, so more agentively: *er, alle, jed, viel, manche, Student, Arbeiter, Frau* ('he', 'all, everyone', 'every, each', 'many', 'some', 'student', 'worker', 'woman'); with an emphasis on a family context in BEL in particular: *Mutter, Großeltern, Junge, Oma* ('mother', 'grandparents', 'boy, (the) young', 'granny'); and, less prototypically, but contextually frequently animate *China, Gesellschaft, Regierung* ('China', 'society', 'government');
  - to assign predicates to generic, inanimate, and abstract entities and concepts: *Zeit, Welt, Situation, Entwicklung, Unterschied, Bedingung, Freiheit* ('time', 'world', 'situation', 'development', 'difference', 'condition', 'freedom'), and actors and concepts in the context of societal development: *Universität, Wohlstand, Arbeit, Chance, Technik, Technologie* ('university', 'prosperity', 'work',

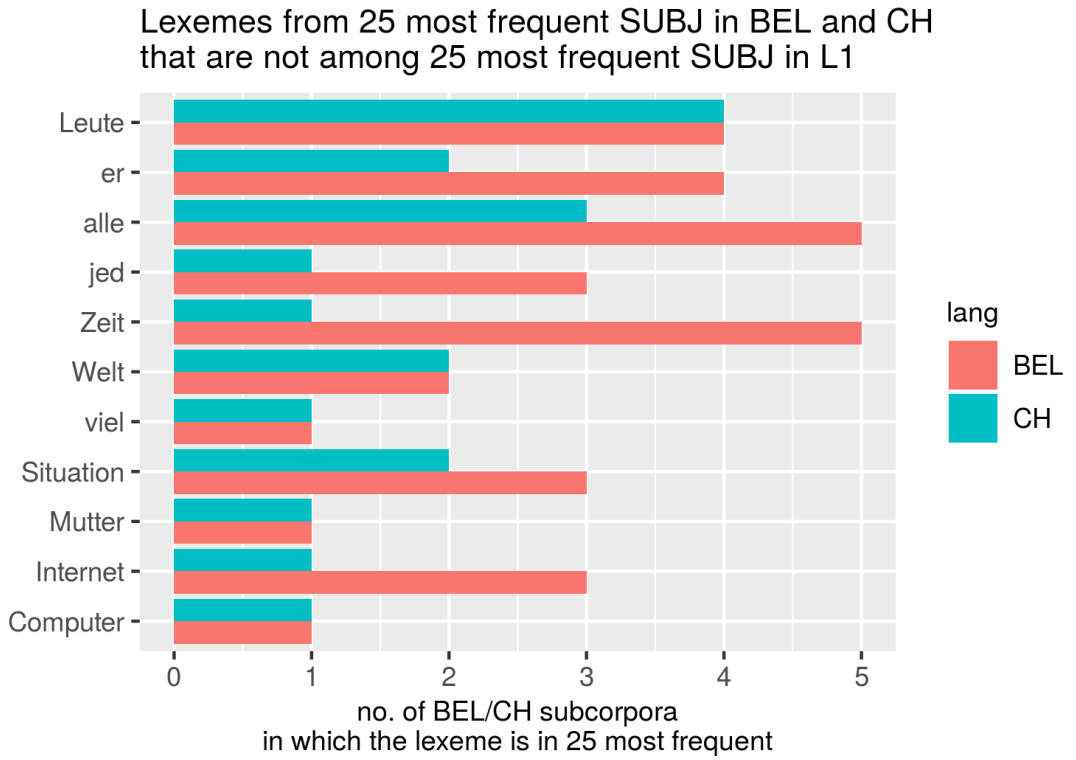


Figure 6.9.: Frequent subjects in L1 but not L2

'chance, opportunity', 'technical equipment, technology', 'technology').

- Grammatical subjects frequent in L2 are introductory rather than anaphoric, except for the two most frequent (*d (der, die, das)* and *was*, which can be used referentially (as demonstrative/anadeictic or relative pronouns, 'which'), but in the case of *was*, also as an interrogative pronoun ('what'). Three of the infrequent grammatical subjects in L2 that are frequent in L1 can be used textanadeictically (*welch, dies, wer*; *welch, wer* also as interrogatively).[11]

It appears then, unlike the predictions in the previous chapter, that subject and verb a) *do* form a functional complex in both L1 and L2 writing, and that b) form and function differ between L1 and L2.[12]

The no_subj graph type is a graph of higher specificity for learners, such as vas_prep is to vas_no_prep, but what is taken away from the vas_no_prep graph is not simply noise in an argument structure analysis. Instead, the difference between no_subj and vas_no_prep appears to represent functional differences in aspects of L2 lexicosyntax. The curious thing here is that it is not only the level of modularity, but also the distribution

[11] Learners use fewer question marks, likely indicating fewer questions in their writing, CH learners especially. However, there are notable stylistic exceptions in both learner groups (8 in BEL, 2 in CH) with rates of >0.01 question mark/token, compared to a median of roughly 0.005 for L1 and BEL-130, BEL-160 and BEL-75 and lower in the other subcorpora.

[12] See next chapter for a discussion of predicates in this context and a possible variationist approach to develop a better understanding of the interaction between form(s) and function in L1 and L2 writing.

that is changed in L1 and L2, suggesting that the subjects although fewer in number, have an equalizing impact on the vas_no_prep graph. Subject lexemes were hypothesized to *add* randomness, not level it; And learners were hypothesized to show clearest effects in their developmental trajectory for the subjectless graph. Both cannot be confirmed here, which is also due to learners showing stark differences with respect to lexical choices and lexical diversity in SUBJ slots compared to OBJA slots. The unexpected results could be explained through the inclusion of pronouns in the analysis, which are often used in SUBJ slots. Taking these connecting hubs out of the graph through the exclusion of subjects could weaken the connectivity of subgraphs where pronouns are particularly frequently used in both SUBJ and OBJA. Whether this is an adequate explanation, and what this implies for the model of coselectional constraint, remains for future research.

Aside from the no_subj graph, hypotheses were confirmed with respect to graph specificity: There is a qualitative difference between the other verb-specific graphs and the full graphs in their levels of modularity, variance, and their trajectories in learners. This suggests that the verb-specific graphs are not just slightly less random full graphs, but that there is an actual division into three structural groups in this model. Since the research question aims at high(est) specificity, but stark differences between two verb-specific graphs were not predicted in previous hypotheses, I will mainly consider the vas_no_prep and the no_subj graphs for further analysis in this chapter.

6.3. Validation

The two most sensitive aspects of this study are corpus size and the grouping of texts into subcorpora. This would be the case regardless of the applied measure, but it is of particular relevance to validate against confounding factors because the mechanics of the measure itself in corpus data are not well-understood yet.

A grouping seems necessary not only to create larger subsets for the analysis, since individual texts are expected to contain insufficient lexicosyntactic material, but also since coselectional constraint itself is an emergent phenomenon: The writing of a single speaker, no matter how long, would not yield information about the coselectional constraints of the community, since coselectional constraints could not be told apart from idiosyncratic preferences.

Splitting a continuous variable such as the onDaF scale into discrete groups (1) holds a risk of misassigning groups, i.e. splitting at linguistically meaningless cut-off points implying they are meaningful, and (2) forces the analysis to artificially separate two values that lie closer to one another while keeping two other values within a single group despite a greater difference. For example, if groups were assigned by onDaF score ranges of 20 starting from zero, texts of scores 119 and 121 would be assigned to separate groups, while 121 and 139 would be in one. This problem is sometimes circumvented by using only data at the center of each target group (e.g. only texts in the onDaF range of 120-125 for *higher-intermediate* and only 140-145 for *advanced* in this analysis). However, Kobalt does not offer sufficiently large groups in sufficiently separable onDaF ranges for this, and even if it did, it would not provide a solution for (1).

Different grouping can imply different corpus sizes if all texts within a certain range are considered in the analysis. But even in balanced groupings, corpus size needs to be investigated independently to gain a better understanding of the mechanics of modularity analysis, including the robustness of trajectories across corpus sizes (are trajectories volatile, more clearly defined, nivellated for larger or smaller corpora?), and the occurrence

of floor and/or ceiling effects. The aim of this section therefore is to clarify,

- whether results from the initial onDaF grouping can be confirmed in other groupings and corpus sizes;
- whether differences in coselectional constraint are detectable in individual documents;
- whether and how linguistic (onDaF-based) grouping and modularity interact;
- how corpus size and modularity interact;
- which corpus size is best suited to capture the effects;
- if any (and if so which) of the corpus sizes analyzed here (1, 5, 6, 10, 15, 20 texts) serve as lower and/or upper bounds for best results,

thereby assessing whether a research question as the one posed at the beginning of this study can reliably be measured in a small to mid-sized corpus such as Kobalt; which aperture provides sharpest results; and whether a simple language assessment like onDaF is helpful as an ordering and grouping variable is helpful to the case.

6.3.1. Corpus size

6.3.1.1. Individual documents

Coselectional preferences are saturation effects, where a category slot has a limited number of potential fillers which all occur eventually, but at different, and changing, frequency rates. It is complementary to the producitivity of verb-argument selection in that a more productive verb-argument slot allows for more *novel* or *different* arguments, while a more constrained verb-argument slot allows only for *specified* arguments. There could be a difference between coselectional constraints and coselectional preferences, whereby coselectional constraints still apply to highly productive verbs and their slots, i.e. that slots allow for qualitatively *specified*, but quantatively *many* arguments; while in coselectional preferences arguments are specified and hence their number is limited. A clear distinction between these two remains for future modeling and research.

Although they have a cognitive extension, too, coselectional constraints in L1 are more strongly a phenomenon of *langue* more than *parole* (Saussure, 1916/1983), which means that they are defined by and most clearly expressed in the language of a community rather than an individual speaker. They are also conceptually related to other phenomena of linguistic convergence and alignment, i.e. sociocognitive processes in dialogue in which speakers' linguistic behavior interactively converges to a mutual frame of reference, shared syntax, and even modulated articulation (Branigan et al., 2000; Steels and Loetzsch, 2006; Pardo, 2006). For coselectional constraint, the process of convergence does not occur interactively in individual dialogue to the same degree, but through language input and enculturation. It is also, in total, much less clearly observable for speakers and linguists alike, because unlike the number of phonemes or different ways of referring to space and time, lexicosyntactic combinations are virtually unlimited (see section 4.2 for a discussion and some example computations).

Coselectional constraint is thus an emergent phenomenon that can only be observed from the comparison of several speakers:[13] If we met a single speaker of a language and studied their coselection of lexical and/or syntactic items, we would be unable to discern between personal preferences and commonalities for frequent coselections, and we would be equally unable to discern between random single occurrences and coselectional constraints. This is reflected in the study of phraseology, where all observations start from the individual word rather than the individual speaker – After all, it is "you shall know a word by the company it keeps"(Firth, 1957, 11), and not "you shall know a word by the various company it keeps depending on who uses it", although of course individual preferences exist as well.

Coselectional constraint can also only be acquired in the context of a group of speakers, because, whether guided by very intricate semantic, semiotic, or morphophonotactic rules or arbitrary and idiosyncratic, its patterns are elusive to a simple description. Unlike the notion of a communicatively efficient *basic learner variety* (Klein and Perdue, 1997), which in theory can be assembled from few linguistic parts without further instruction, coselectional constraint apparently requires a large amount of contextualized input for acquisition (see section 2.2).

Yet, it is still the cumulated *individual speakers' production* that reflects the fact of coselectional constraint. Coselectional preferences of a language community are simultaneously created through the production of individual speakers and, in turn, shape it. They can only be observed relative to the group, and are also flexible with respect to individual variance. Coselectional constraint thus is an emergent and synergistic phenomenon, i.e. one where the individuals and their belonging to a group interact and create the phenomenon interdepenently. In this sense, coselectional properties as lexicosyntactic phenomena are difficult to observe in individual production of language such as a single essay.

At the same time, if abstracting from concrete items, coselectional constraint is also a supraconstructional or structural property of the lexicon and lexicosyntax. Assuming that there is something like structurally diverse interlanguage depending on the acquisition stage (Selinker, 1972), and that this interlanguage is to some degree independent of the individual learner – no learner covers all of their according interlanguage space, but all share some common ground – structural differences in lexicosyntax should exist between learners. Since any text can be modeled as the instantiation of a lexicon, its structure is measurable even in individual documents with the obvious limitation of small corpus size.

Figs. 6.10 and 6.11 show that modularity is overall higher in individual texts than in the grouped graphs. The regression ranks graphs by specificity, beginning with the least specific full graph, though at much higher levels than in the grouped analysis (lowest modularity in full graphs in onDaF-based groups was 0.3, now 0.41). However, variance is so high that data points of graph types that were 0.2 or more apart in modularity in the grouped plots are now overlapping, and some of this overlap can likely be attributed to a ceiling effect (modularity reaches values > 0.85 in all three language groups, maximum defined modularity $= 1$). At the same time, it reflects the reality of the Zipf-distribution of lexical items, where a graph becomes more sophisticated with each new word that is not included in the previous text, while the most frequent words will appear in all texts without adding new nodes to the graph. This means that shared vocabulary will weigh modularity down in all corpus sizes larger than a single text.

Figs. 6.10 and 6.11 also show that, indeed, a trajectory with a drop towards intermediate

[13] In the same way that alignment can only be captured as a dynamic process in dialogue. Structures emerged from alignment may last, but would be unobservable unless a pre-aligned state is somehow captured for comparison.

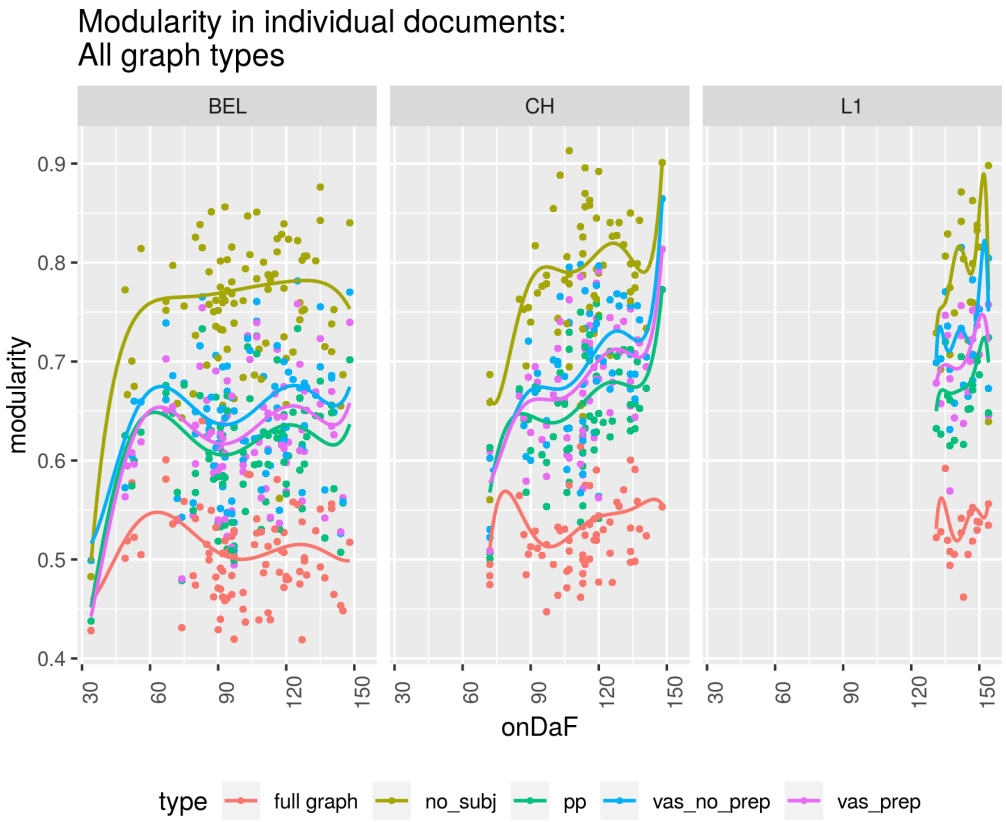


Figure 6.10.: Modularity of graphs derived from individual documents. A u-shaped development is visible in the verb-specific graphs in BEL in verb-specific graphs, except no_subj; and possibly hinted at in CH. Early data in BEL (30-60 onDaF points) is rather sparse and texts are rather short, thus the beginning of the trajectory should not be overrated. On the other hand, CH also starts from a lower modularity at around 70 onDaF points before rising and dropping at around 90 in some graph types. This suggests that there might be another process at play in this onDaF range.

stages is visible even for individual documents in BEL (except no_subj) and is hinted at in the CH-data as well. Modularity is growing in a steeper sloper in CH vs. BEL for learners scoring about 120 onDaF points and higher, and reaches higher final values even than L1. Trajectories are more similar between the three verb-specific graphs without no_subj than full graph and no_subj.-

A slight negative text length effect is observed for vas_no_prep in L1 and L2, and for no_subj in L1, but not L2 (fig. 6.12). The effect in L1 is also stronger for no_subj than vas_no_prep. This corroborates the interpretation that Louvain modularity indeed works as a measure of coselectional constraint: Consistent with the results presented in section 6.2, taking out subject lexemes in no_subj in L1 raises network redundancy. Since subjects are introduced more frequently compared to accusative objects, graphs without subject lexemes are less structured because they lack the hapaxes from subject introduction. Since accusative objects are saturated more quickly, the effect stronger for longer text. This can be seen as a first indication of construct validity.

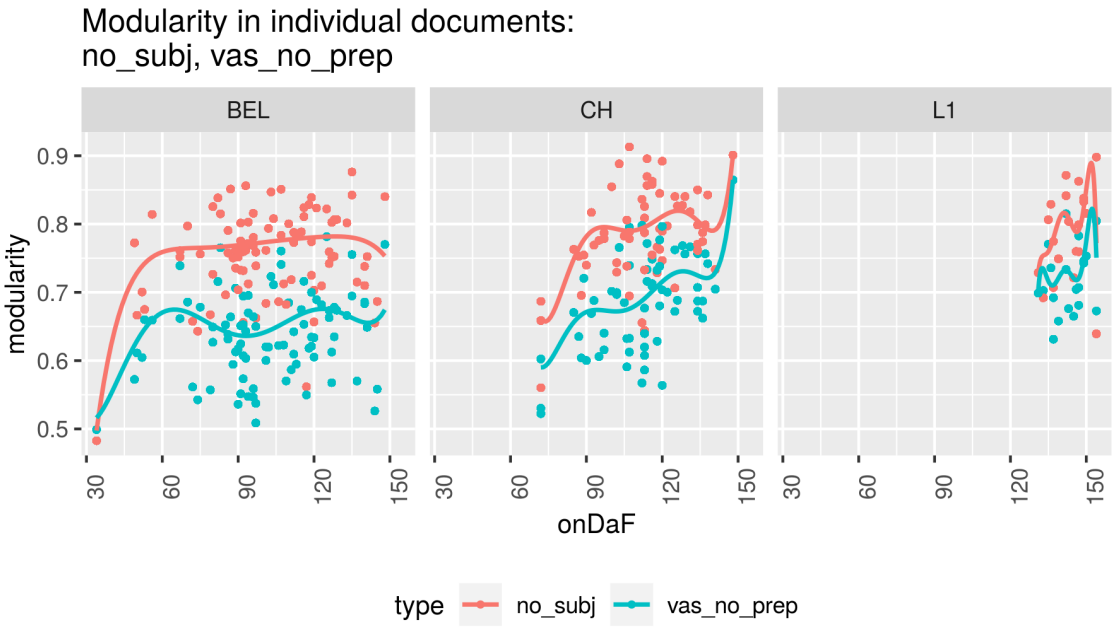


Figure 6.11.: Modularity of no_subj and vas_no_prep graphs derived from individual documents

The observations in this section also serve as a first implicit validation of the onDaF-based grouping: Trajectories, as they were implied in the grouped analysis, are continuous in the individual text analysis. Box medians, as they were presented in fig. 6.1, do represent relevant aspects of the trajectory as it exists in this analysis, too. An interesting difference is the apparent absence of a u-shaped curve in BEL learners in no_subj. This could be an effect from the genericity of subject lexemes used by learners: Perhaps in a single corpus, they are used rarely, such that the graph is not less structured, but in the corpus, they are repetitive, causing a drop in modularity. This would be direct evidence for the dialectics of emergent vs. individual effects. It cannot be confirmed with certainty though, since variance is so high that the regression can only be seen as a vague representation of an implied trajectory.

With the clear ranking by graph specificities and the overall agreement with predicted behavior (u-shaped-trajectories, lower modularity for learners vs. L1 (except CH $>$ 120 onDaF points), it appears that modularity values do not drift into randomness even in corpora as small as several hundred tokens. Rather, the analysis suggests that lexicosyntactic constraint is to a certain extent quantifiable even in individual documents. This is a pleasantly surprising finding considering that in the communities around DH and corpus linguistics, the call to collect more data and build larger corpora has been strong, while small datasets are often reported in absolute numbers or with statistics of limited power. It appears from this analysis that graph metrics might provide enough additional information to the analysis (compared to a reduction to lexical combinations taken from texts) to gain insight even in very limited data.

However, at the same time, with the large variance and overlap and an unclear role of floor and ceiling effects, it does not look like the single text perspective is ideally suited to capture the predicted effects with great clarity, as was expected in arguing for idiomaticity as an emergent phenomenon.

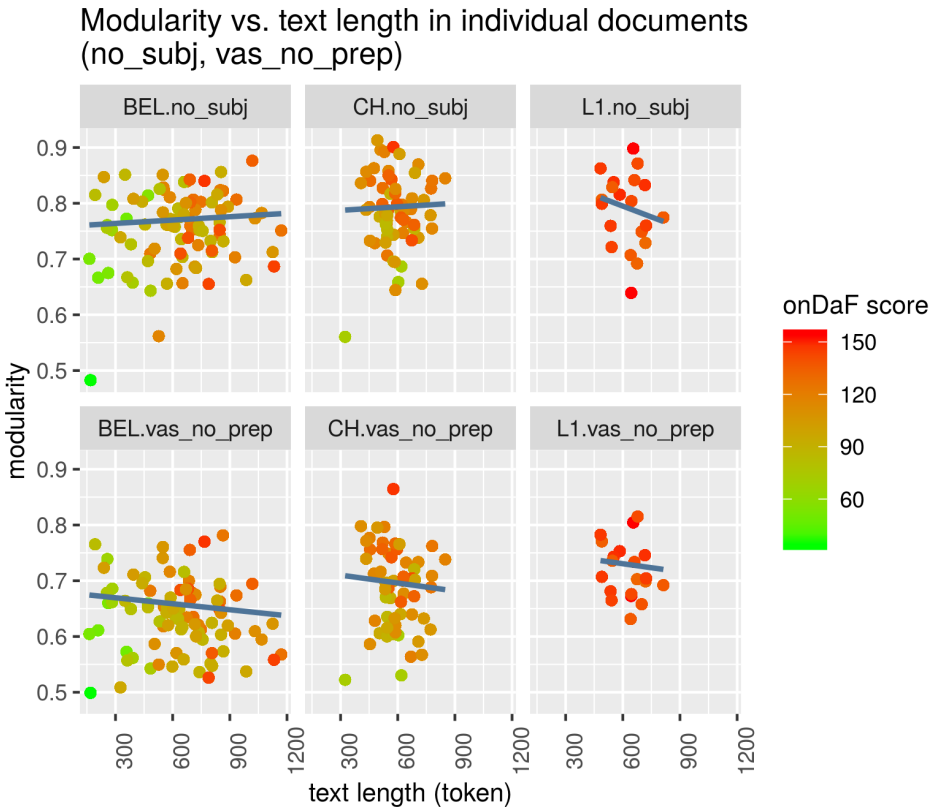


Figure 6.12.: Modularity of no_subj and vas_no_prep graphs derived from individual documents vs. text length. Different trajectories in vas_no_prep and and no_subj are more clearly defined than in the grouped analysis. This could be effect from the genericity of subject lexemes used by learners.

6.3.1.2. Smaller onDaF ranges (onDaF10)

One debatable aspect of the earlier onDaF-based grouping is that it groups data relatively far apart in onDaF scores (15-30 points) into the same group, and the low number of data points, i.e. subcorpora, this provides for comparison (four and five in the BEL- and the CH-data respectively). This means that a) it is possible that the grouping smooths over existing differences between the higher and the lower end of each group, therefore covering up an existing u-shape in the CH data, and that b) a lower modularity value between two higher ones suggests a u-shape, but cannot be told apart from an outlier, such as in the BEL-95 no_subj between the higher neighboring groups (see fig. 6.16, reproduced here for easier comparison, but otherwise identical to fig. 6.3). A more fine-grained analysis of the trajectory at the critical intermediate stages is desirable to assess the continuity of the u-shaped curve in BEL (rather than a w- or M-shaped development or simply erratic behavior) and to gain a better understanding of the trajectory in CH, too.

Splitting Kobalt into smaller subcorpora based on 10 onDaF points can only be done idiosyncratic cut-off points (74, 84, 94, etc.) if the number of intermediate groups and the minimal number of texts in each group are to be maximized. With this, total numbers of texts in each group vary greatly: If all texts in a group were included in the analysis, the largest corpus would reach a size of 18 texts (6th group in BEL), while the smallest only contains 6 (10th group in CH, see fig. 6.13). Again, the smallest groups are those

at the edges of the grouping, systematically skewing results towards a u-shape. To avoid this, samples of six texts are compared in figs. 6.14 and 6.15. All native speakers were assigned to a single group, even though they are technically split almost evenly into two groups (10 L1 in group 12, 7 in 11) and some scored even lower than that (3 in group 10).

OnDaF groups 1-5 and 12 for the learners were excluded in this analysis due to low number of texts, leaving 6 groups for comparison (+2 for CH, +1 for BEL). However, rather than clarifying or strengthening results from the previous analysis, the figs. 6.14 and 6.15 show unclear trajectories at high variance and overlap between graph specificities, such as equal modularity medians in L1 for vas_no_prep and vas_prep in L1.

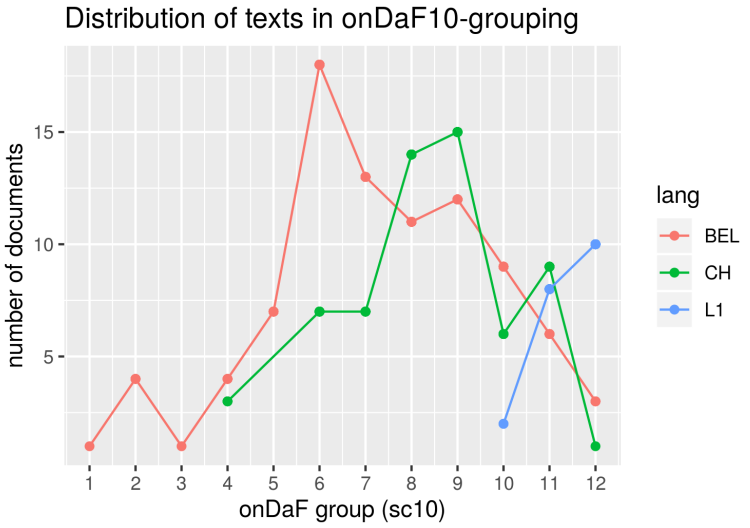


Figure 6.13.: Number of documents in onDaF10-based subcorpora.

It is possible that the initial onDaF-based-grouping does actually reflect acquisition stages in a way that a grouping based on smaller onDaF ranges cannot, and that the cut-off points, though not psychometrically validated here, do coincide with linguistically meaningful points of transition; or that the balanced grouping simply reflects a good division of the data as it is distributed. This will be discussed further in the sliding window analysis in section 6.3.2.2. More likely, however, six texts is not a good size for a valid comparison. Outliers may have too great of an impact in a corpus size as small as this in leveraging the modularity of the whole corpus up and down to an unnatural degree for the respective acquisition stage. This would mean that the grouping fails because individual effects are larger than group effects. The following section looks into this in an out-of sampling.

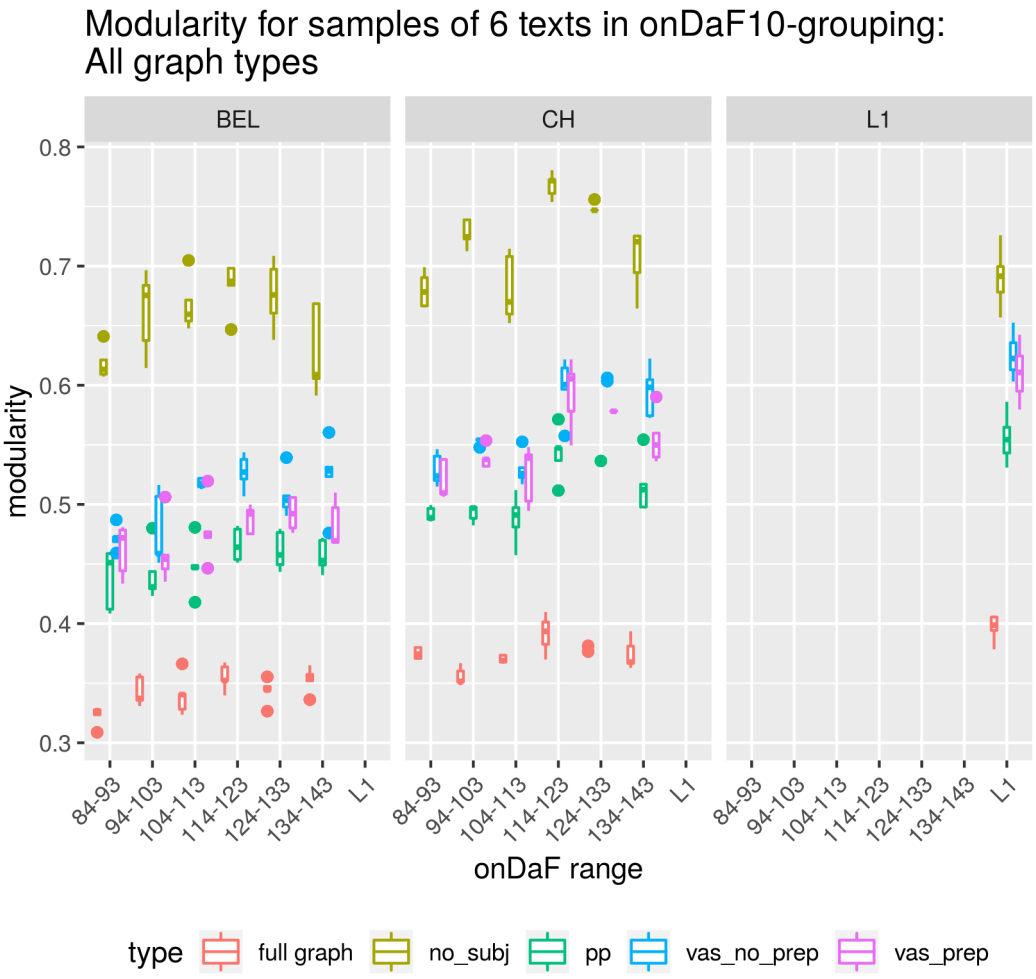


Figure 6.14.: Modularity for onDaF10-based subcorpora, five 6-text-samples per box

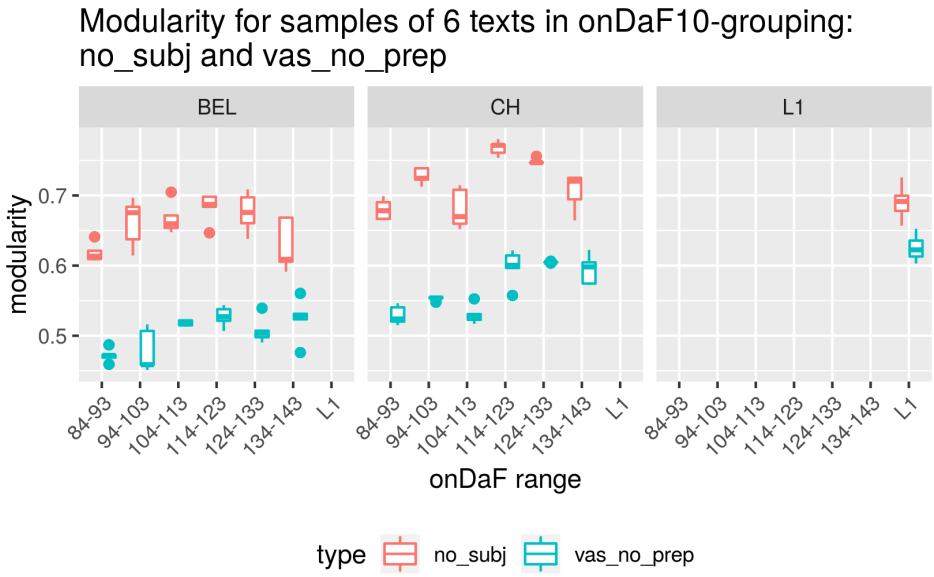


Figure 6.15.: Modularity in onDaF10-based subcorpora, five 6-text-samples per box

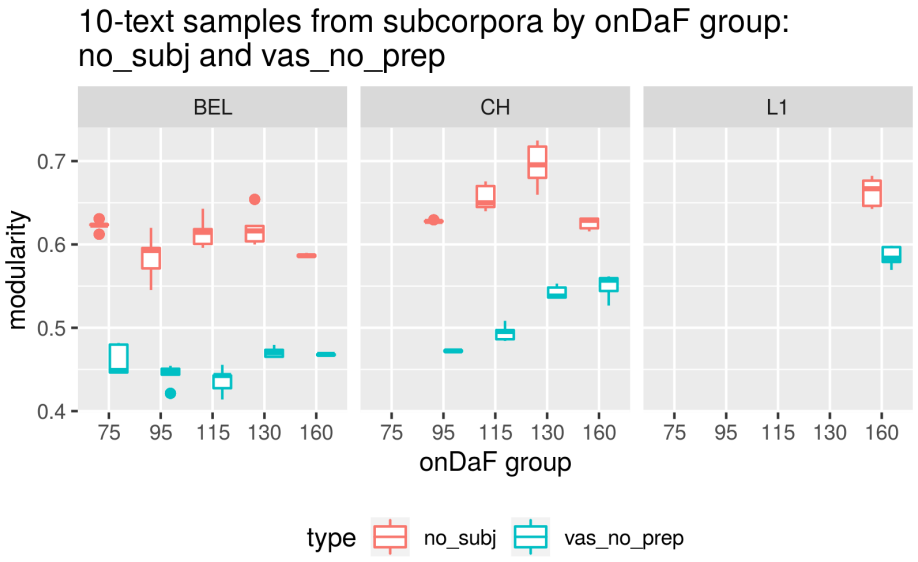


Figure 6.16.: Modularity in 10-text samples from earlier onDaF grouping, 5 samples per subcorpus, reproduced here for comparison, identical to fig. 6.3

6.3.2. Grouping

Grouping and corpus size cannot be isolated in Kobalt due to the distribution of texts – a grouping by smaller onDaF-ranges in this dataset implies smaller corpus size. This section will therefore address both aspects. First, a larger group vs. individual effects for the initial grouping, but not the onDaF10-grouping are shown. Then, a sliding-window-sampling is performed to show that modularity values align into rather neat trajectories by onDaF median for window sizes of 10 and more texts, where and that the initial onDaF-grouping is not inferior to an analysis based on a simulated continuity of data points rather than a grouping.

6.3.2.1. 5/6- and 9/10-sampling

To gain more insight into the role of individual variance, an out-of-sampling has been performed on the two groupings (onDaF and onDaf10, 9/10-(9-out-of-10)- and 5/6-(5-out-of-6)-sampling respectively). This sampling technique is a simplification of k-fold cross-validation and leave-one-out cross-validation, which is a technique used widely in machine learning to validate the performance of an algorithm trained k-1 splits of the data and tested on the last split in all permutations. Here, it is used to estimate the impact of individual variance on the grouped corpora and their modularity values.

For each of the five 6-text-samples in the onDaF10-splits, one text was left out, and modularity computed on the remaining five texts. That way, each sample was sampled 6 times, leaving out a different text in each sample:

- Samples 1 through 6 include different texts, they are sampled from the corpus.
- Samples 1-1 through 1-6 include the same texts with one left out each: Sample 1-1 contains texts 2, 3, 4, 5, and 6 from sample 1. Sample 1-2 contains texts 1, 3, 4, 5, and 6 from sample 1, and so on.
- The same was done with the 10-text-samples from the previous onDaF-grouping, where sample 1-1 contained texts 2-10, sample 1-2 contained texts 1 and 3-10, etc.

Thus, each of the 5/6-samples makes a corpus of the size of 5 texts, while each of the 9/10-samples makes a corpus of the size of 9 texts. Fig. 6.17 shows a comparison of the two sets of samples in L1 with a detailed explanation of how to read the plot in the caption.[14]

In L1, corpus size effects in vas_no_prep are consistently stronger than individual effects. In no_subj, this is less clearly defined, but still true of most samples. In all learner graphs, the 9/10-sampling shows larger group- vs. individual effects (blue boxes span different onDaF ranges by group, little overlap between grids), while the 5/6-sampling shows large variance and stronger effects from sample no., i.e. texts chosen for analyis, particularly at intermediate stages (this is despite equal corpus sizes in the sampling). This supports the conjecture that a corpus size of five texts is not well-suited for an analysis of this data, and is in fact less telling than either the individual text analysis or the 10-text-samples from the initial analysis. It also shows that variance in learners is highest at intermediate stages, which is consistent with the prediction of a process of randomization

[14] These plots are little tricky to read, but I believe it is worth bearing with them, because they provide a way of assessing the validity of a grouping vs. individual effects, which is difficult without abundant data and a recurrent issue in corpus linguistics.

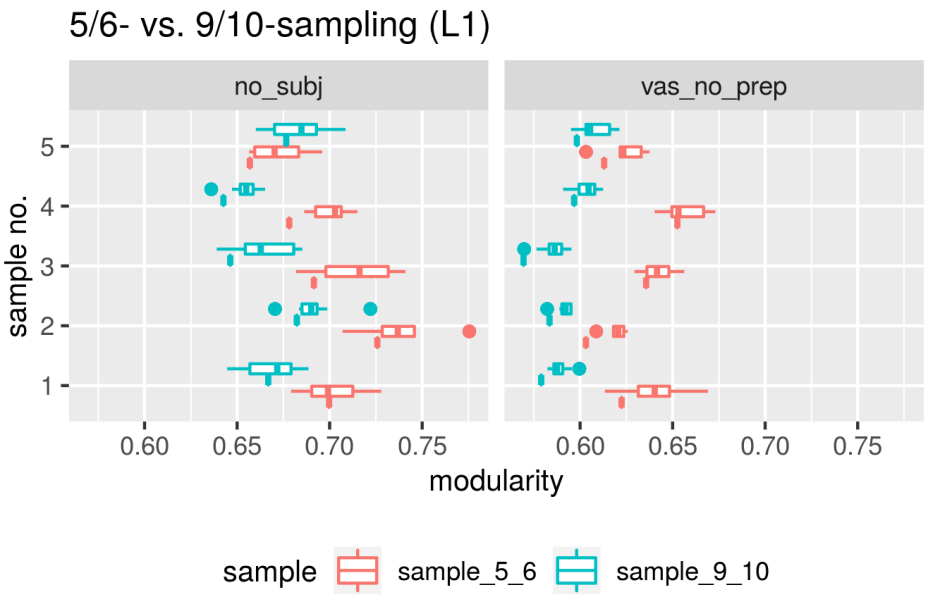


Figure 6.17.: Comparison of 5/6- and 9/10-sampling in L1. The plot is read as follows: The red boxes represent sampled 5/6 samples, i.e. sample 1-1 through 1-6 in the lowest row. Samples 1 through 6 include different texts, they are sampled from the corpus. Samples 1-1 through 1-6 include the same texts with one left out each. The little red line underneath the box represents the full 6-text-sample. Larger corpus size should lead to lower modularity. Thus, boxes should all drift away *rightward* from the little line underneath. If individual effects are larger than corpus size effects, boxes hover above the full sample line. Red boxes should lie rightward of blue boxes, since they come from smaller corpora. In vas_no_prep, corpus size effects are larger than individual effects in the 9/10-sampling in all samples except sample 4 (box covers range of line). In no_subj, results are mixed, but except for sample 1, all 9/10-samples are less modular than the 5/6, i.e. the corpus size effect is larger than individual effects.

(since randomization would be expected to happen, well, randomly for individual learners rather than for the whole cohort at once).

Perhaps 5 or 6 texts are a size at which the beginning emergence of communal properties and individual preferences interfere to an extent that can, but does not necessarily break an existing pattern. This is further corroborated by the variance plot in fig. 6.22, where in the vas_no_prep graph, variance between 9-text-samples decreases with increasing onDaF group in BEL and is low across groups in CH, while variance in 5-text-samples behaves seemingly randomly and reaches much higher values. In the no_subj graph, variance is inversely u-shaped in both language group in the 9/10-samples, but erratic in 5/6-samples.

While not constituting a lower bound (because a regression over individual texts seems to yield better results), it appears that the onDaF10-grouping spans a low turning point at which individual variance and emergence interfere in a way that confounds the analysis. It also seems that simply splitting the onDaF scale into smaller units by itself does not

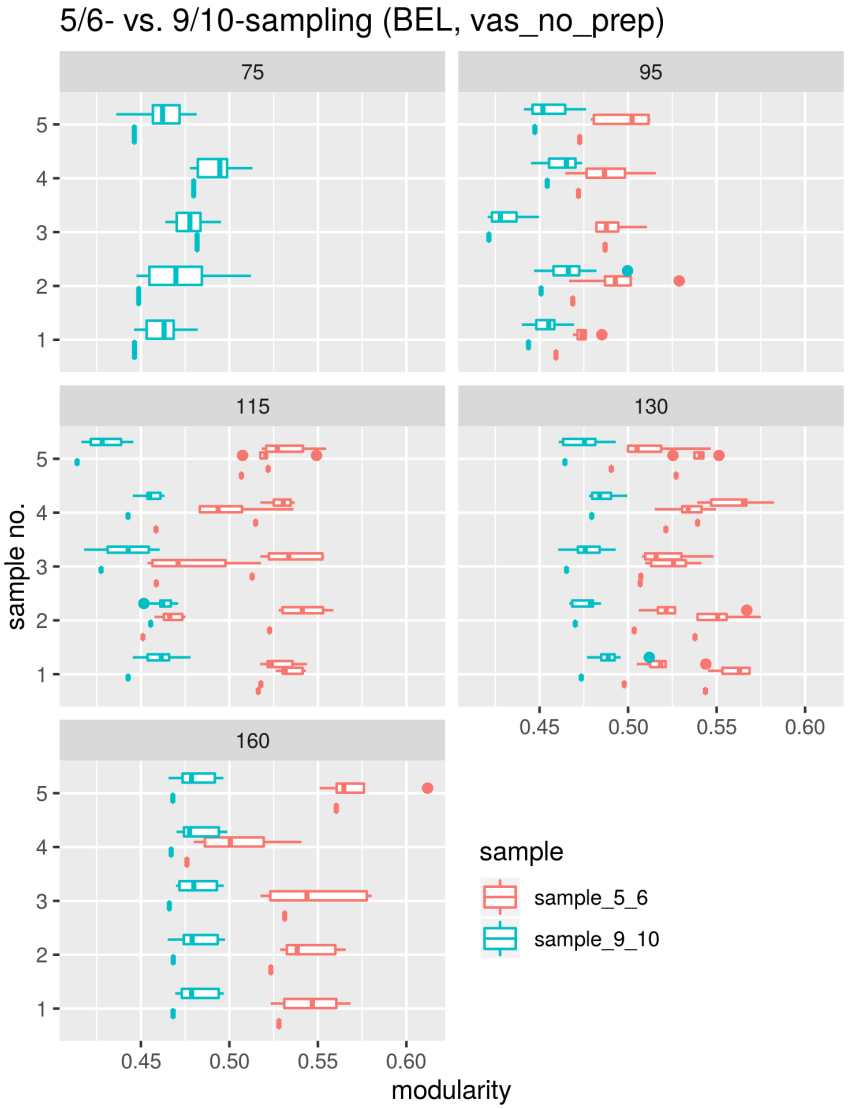


Figure 6.18.: Comparison of 5/6- vs. 9/10-sampling in BEL, vas_no_prep. Numbers in grid headers reference the initial onDaF groups, where in 115 and 130, there are two onDaF10-groups. In 75, there was no group of 6 texts within the span of 10 onDaF points, thus there is no 5/6-sampling for this group. Sample boxes should shift rightward of the full sample line underneath, which is largely true of the 9/10-samples, but not of all 5/6-samples (consider sample 3 in 95, 3 and 5 in 160). In addition, all samples in a grid should cover the same modularity range, if samples were similar to one another in modularity values (i.e. if the effect of grouping is stronger than the effect from choice of individual texts). This is not the case in 115, 130, and 160 in the 5/6-sampling, but it is more so in the same grids in the 9/10-sampling. This suggests that in a 9-text-corpus, or in the larger onDaF-range grouping (these effects cannot be isolated in Kobalt) is superior to a 6-text corpus in the analysis.

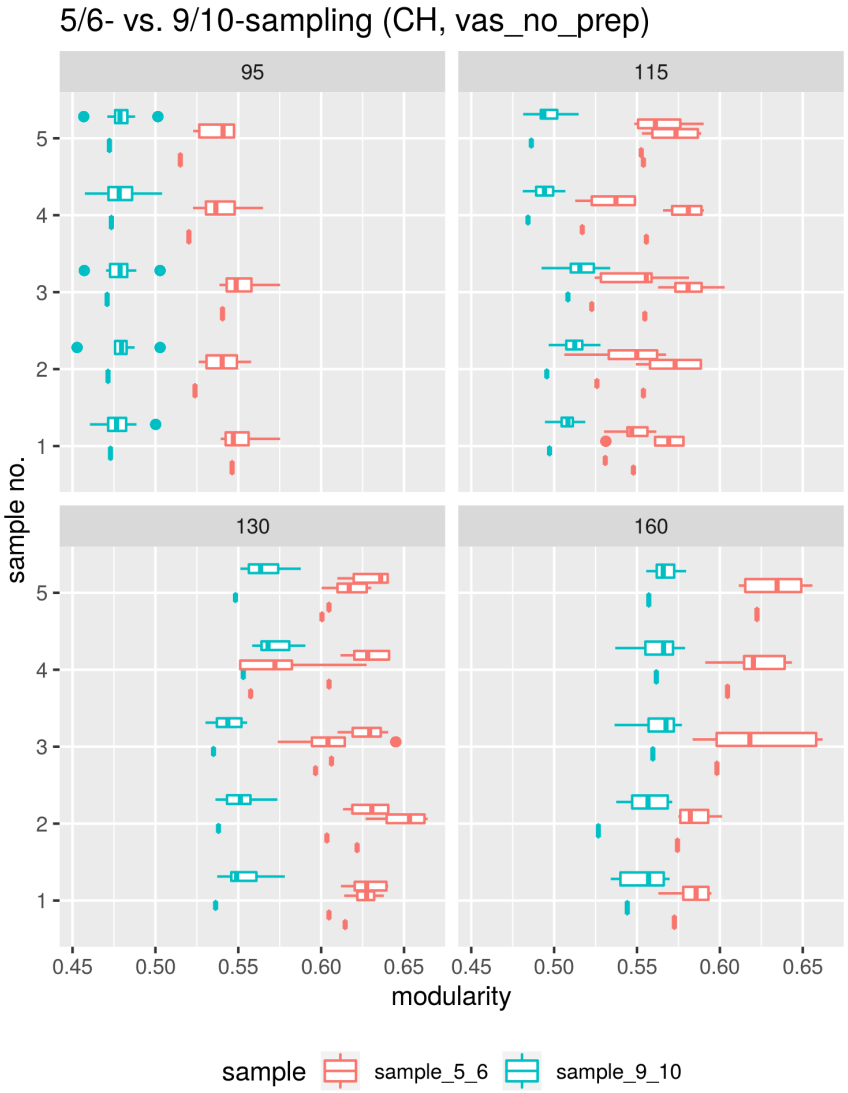


Figure 6.19.: Comparison of 5/6- and 9/10-sampling in CH, vas_no_prep. Both samplings work fine in CH-95, but variance grows in the intermediate learners. It does not reach BEL-like levels though. This suggests that some process of randomization happens in CH learners at intermediate stages too, but less so than in BEL-learners.

create valuable new groups for comparison. This is relevant for the further systematization and development of methods in corpus linguistics, where small to medium-sized corpora are concerned, because it shows that a quantitative analysis of individual texts can yield better results than an analysis of several texts. This is not a trivial in corpus linguistics, and certainly not in the analysis of lexicosyntax which due to lexical distributions seems to require grouped data (as was argued in this chapter, too). It appears that in some cases, smaller data is less confounded by emerging factor interactions than grouped data.

The out-of-sampling has confirmed that the initial grouping yields group results rather than conflations from individual variance. A division of groups into onDaF10-groups on

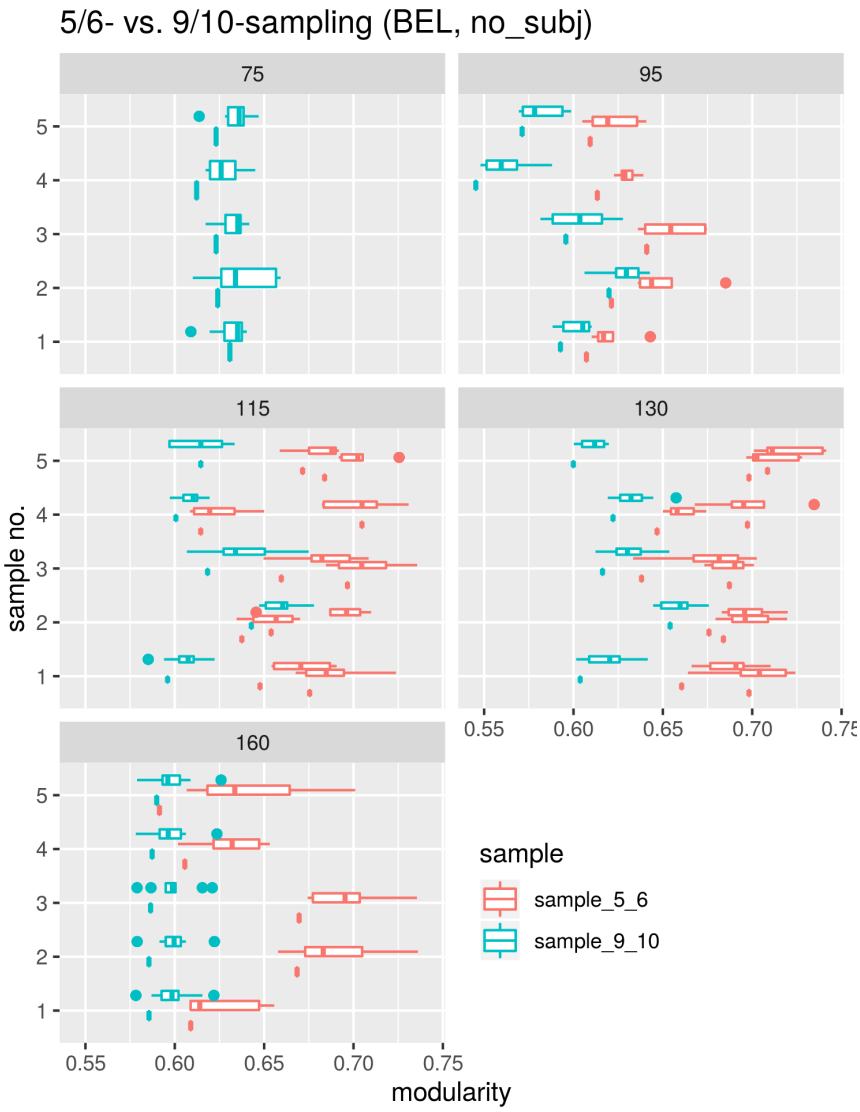


Figure 6.20.: Comparison of 5/6- and 9/10-sampling in BEL no_subj. 9/10-sampled graphs do not align as neatly as in vas_no_prep, suggesting variance is overall higher. This could be partially attributed to smaller corpus size (since subject lexemes are removed from the graph, it is smaller than vas_no_prep). But blue boxes here are still less volatile than red boxes in BEL vas_no_prep, suggesting that variance is also generally higher in no_subj. This was also found in the other analyses.

the other hand was unable to answer the question posed at the beginning of this section, whether a more fine-grained analysis at intermediate stages reveals more clearly defined trajectories, and it is still unclear how modularity interacts with corpus size in detail. This is why the next step in validation of corpus size and grouping choices is to give up fixed onDaF ranges and group by fixed corpus size and onDaF rank instead, effectively simulating a continuous trajectory. With this, corpus size can be manipulated, which makes it possible to compare larger to smaller windows for overlapping onDaF ranges directly.

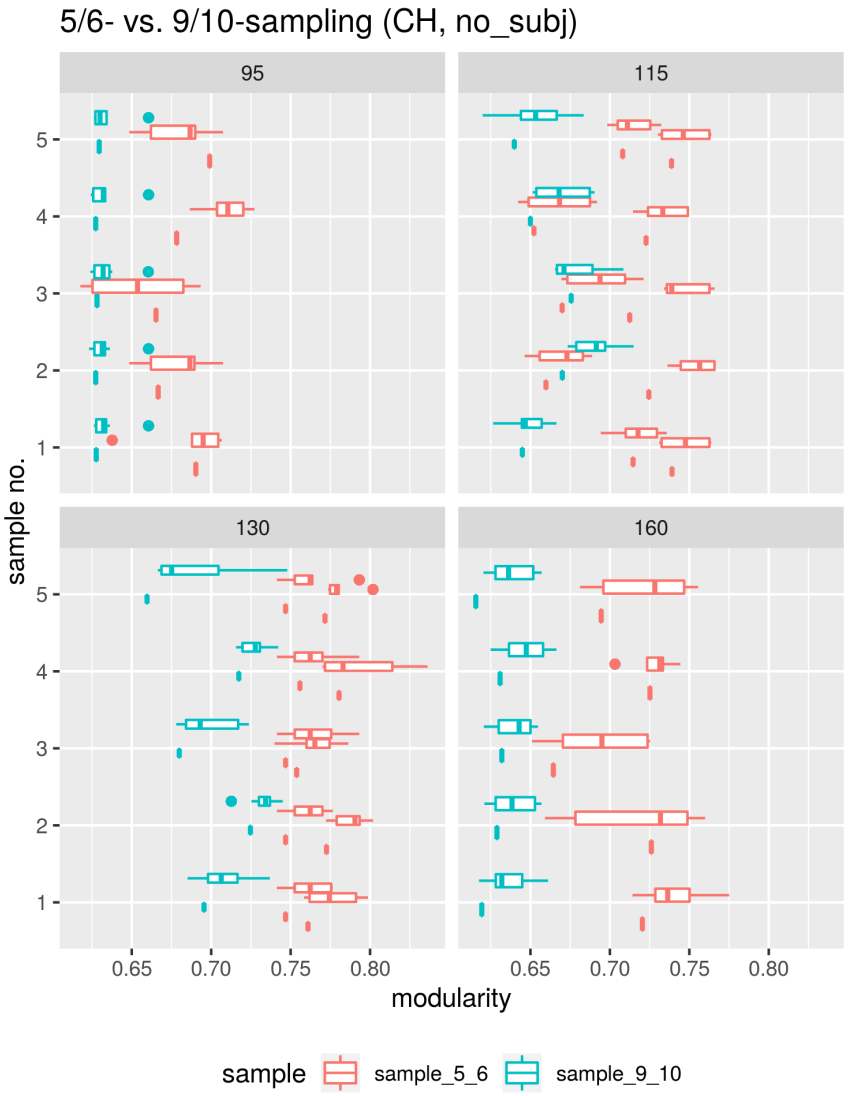


Figure 6.21.: Comparison of 5/6- and 9/10-sampling in CH no_subj. The 5/6-sampling works well for 130 and 160, but not the lower groups. Individual effects are so strong that in 115, two 5-text-samples have *higher* modularity than two 9-text-samples, suggesting their respective graphs are hyperconnected for their size. Hyperconnected graphs are sometimes referred to as *hairball graphs*.

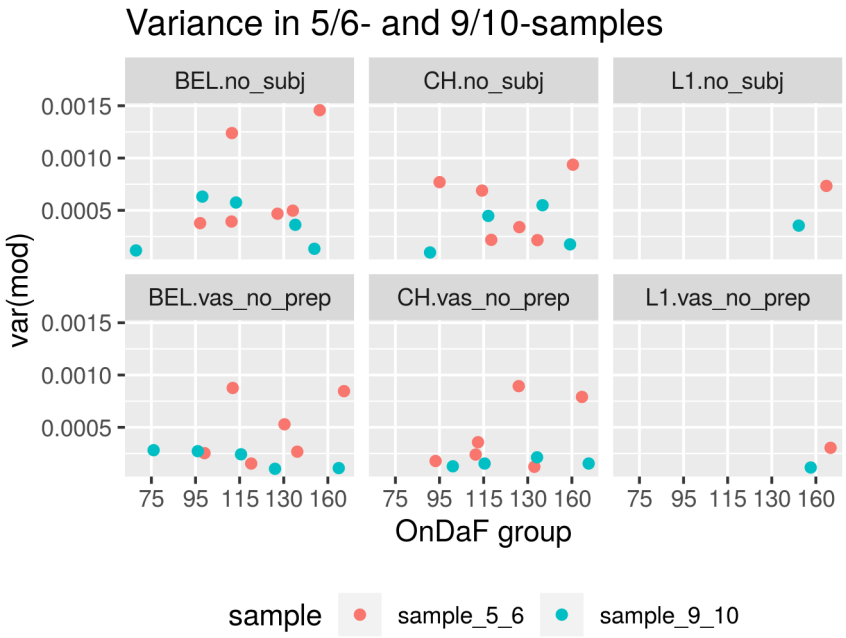


Figure 6.22.: Variance of modularity in 5/6- and 9/10-samples. Variance is low in vas_no_prep in the 9/10-, but not the 5/6-sampling, and erratic in both graph types in learners. Variance in the no_subj graphs is inversely u-shaped in the 9/10-sampling.

6.3.2.2. Sliding window of 5, 10, 15, and 20 texts

While a c-test like onDaF provides limited information about target language skills, certainly productive ones, it is by design correlated with skills measured in more sophisticated testing, effectively triangulating the collected essays to a linguistic model of acquisition stages.[15]

Since the research question implies a model of progression within an interlanguage space with discernable properties based on location within that space, i.e. acquisition stages or skill levels, the previously discussed groupings selected texts by onDaF limitations were related to those by design (Kobalt was intended as a B2-Korpus, see Zinsmeister et al. (2012)), although it was not claimed as a theoretical prerequisite for grouping.

Another way to divide the data into subcorpora is by dropping such a linguistic model and instead grouping by closest ranking neighbors at stable corpus sizes. In what follows, modularity values for sliding windows of 5, 10, 15, and 20 texts neighboring in rank were computed, so in a 5-text-window, texts 1–5, 2–6, 3–7 (...), n_{texts}–$m_{windowsize}$ through n_{texts}. The score range in this approach is random and fluctuating, meaning that the grouping itself does not triangulate to CEFR levels, but the ranking is still linguistically motivated. This is more resembling of a dynamic or continuous model of language acquisition or trajectory through a common interlanguage space than modelling skills as belonging to discrete levels or stages. However, if there are quality leaps at certain onDaF cut-off points, the dense analysis should be able to capture those to a limited extent (they would likely partially be levelled through the combination of lower- and higher-ranking texts in the larger windows). Of course in this dataset, ideal continuity is not reached since even a sliding window model cannot fill the gaps that exist at the higher ranges in both groups and the lower ranges in the CH-group in particular, and the assumption of a common trajectory along onDaF scores shared by diverse learners would be an oversimplification if not aware of other interfering factors.

Sliding (or moving) window sampling is a technique typically used for sampling output of dynamic processes or estimating consistency in signal processing, data streaming or sensor monitoring (Jain and Chang, 2004; Cormode et al., 2010), but also for sampling of non-dynamic entities, such as DNA sequences (Cummings et al., 1995) or acidic rainfall effect on soil quality (Haas, 1990). Recently, there have also been two applications in NLP, where sliding windows of five and ten windows are used for the derivation of a syntactic complexity contour of a text based on a number of syntactic complexity measures. This

[15] Whether that is empirically valid or not. Eckes and Grotjahn (2006) report high correlation rates (> 0.7, p. 311f.) between TestDaF skill assessment (reading, writing, listening, speaking) and onDaF scores, and conclude that the onDaF does indeed measure general language skill. Eckes (2010) gives reliability scores (Cronbach's α) >0.95 for a separability into 6.5–8.8 significantly different groups in test runs with several hundred participants. This provides evidence towards the test's validity at measuring progress, but it does not fully answer the question of how similar or how different two learners' language skills are at identical onDaF scores, and how that relates to certain aspects of their writing (which is not its purpose either, as Eckes (2010, 127) points out). At the same time, modeling "general language skill" is not trivial either, and, as Wisniewski (2017a, 245) notes, "[t]he CEFR levels are not claimed to correspond to a developmental hierarchy in an SLA sense" and the level descriptions do not typically fit the reality of learner language (Wisniewski, 2017a,b). Mapping those four concepts (acquisition stages in SLA, general language skill, CEFR-levels, and measurement through onDaF) to one other is a complex endeavor that lies beyond the scope of this study. However, despite conceptual vagueness of which grammatical, lexical, or processing phenomena are affected specifically and in which way, everyone seems to agree that higher test scores reliably indicate higher skill of some sort, and that higher skill is generally measurable, and many imply quality leaps in several dimensions (lexical, grammatical, skill-specific) rather than a fully continuous development. See also the discussion in chapter 3.2.1.

can then be used to show differences between learner and native speaker text (Ströbel et al., 2016) or for text genre classification (Ströbel et al., 2018). Sliding windows are typically used where processes are dynamic in space or time, or as Braverman et al. (2009, 147) put it:

> "There are two equally important types of the sliding windows model – windows with fixed size, (e.g., where items arrive one at a time, and only the most recent n items remain active for some fixed paramter n), and bursty windows (e.g., where many items can arrive in "bursts"at a single step and where only items from the last t steps remain active, again for some fixed parameter t)."

The Kobalt corpus itself is, if we consider the relation between number of texts and onDaF scores, a single bursty window: For each +1 in onDaF scores, most often no additional text would be included in the graph, while for some scores, several texts would be added, leading to unequal corpus sizes. In this sense, the above subcorpora can be considered a subset of sliding windows of fixed onDaF scores with discarded overlaps. A fixed corpus size of 5, 10, 15 or 20 texts then necessarily leads to unequal score ranges, with the largest windows at the onDaF score range edges having very high ranges of 52 (BEL, first 20-text-window), while some of the intermediate windows of 5 texts, in the bursty parts, differ by only one point (fig. 6.23, in a balanced dataset, the line would be more or less flat at a fixed range, or oscillating between a small number of ranges). Most windows, however, are within a range of 20 points, which is about as wide as the initial onDaF-grouped analysis.

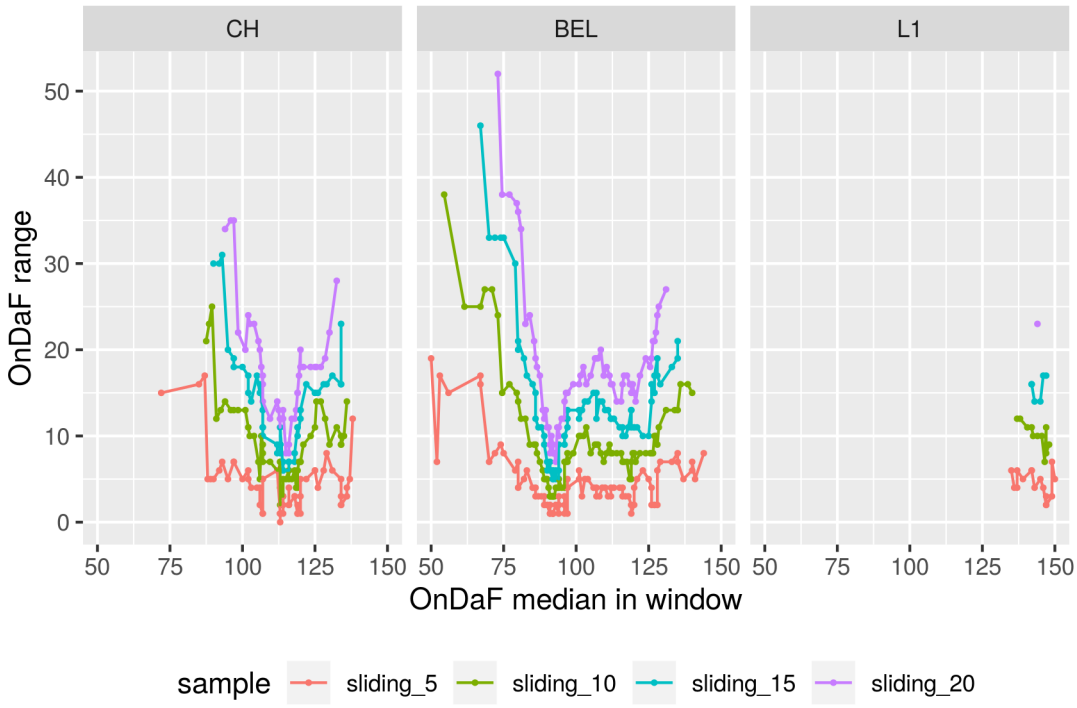


Figure 6.23.: onDaF score ranges in sliding windows

Figs. 6.24 and 6.25 show the distribution of modularity across sample sizes. Corpus size and modularity interact strongly, and more strongly in L2 than L1, but clear patterns

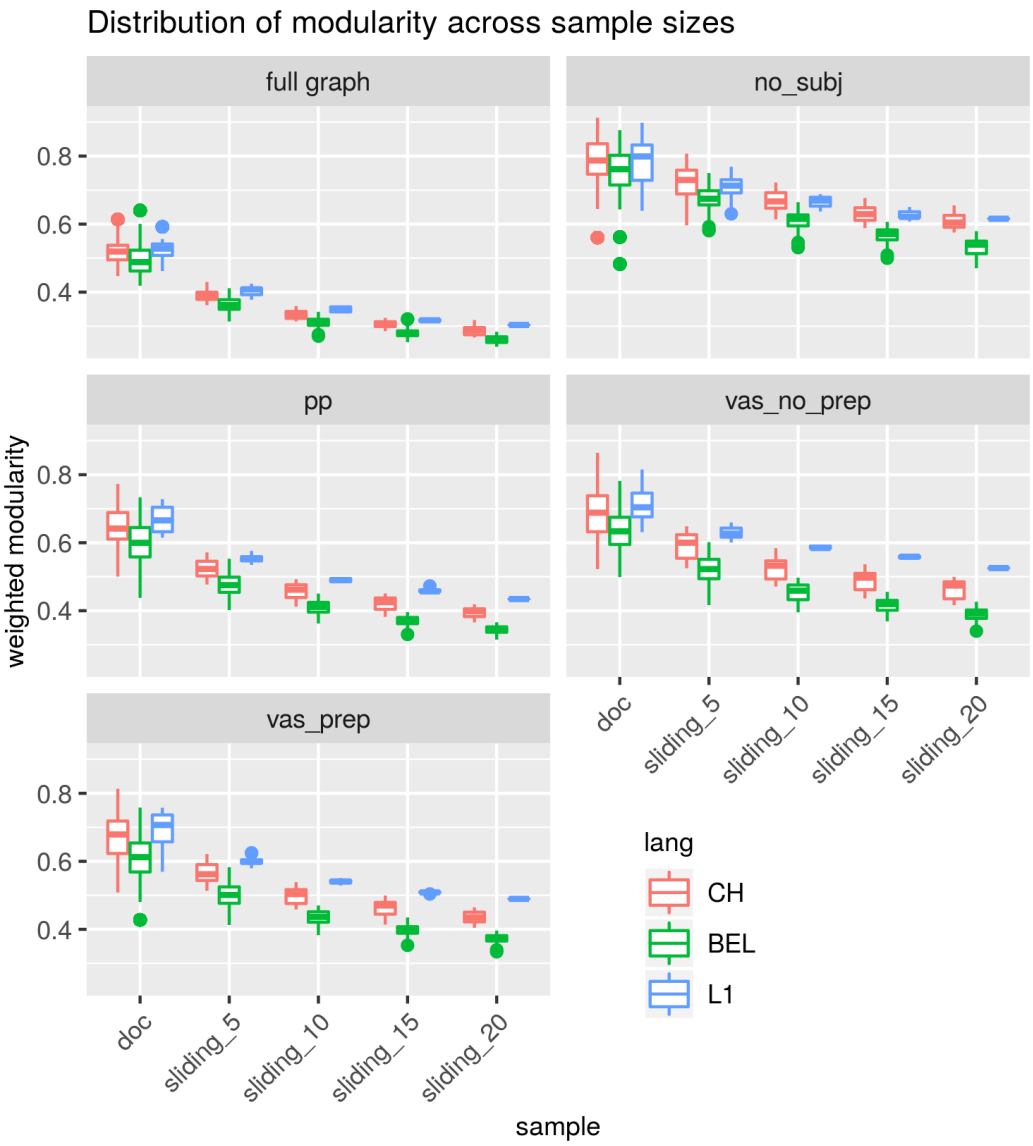


Figure 6.24.: Modularity vs. sample size

emerge early and are stable across sizes. The differences between the three language groups that were observed in the other analyses are confirmed here as a very stable and clear effect across specificities and sample sizes. Modularity is higher at each sample size for L1 than CH than BEL with some overlap mostly between L1 and CH and CH and BEL, but not typically BEL and L1 except for individual documents, and some of the full and no_subj graphs in grouped corpora.

Fig. 6.25 shows that the median nearly converges in L1 with the 20-text window for full graphs, pp, and vas_prep. In L2, convergence is likely reached a little later as extrapolated from the verb-specific curves – it seems, less text is required to reach structural stability in L1, and stability will be reached at more sophisticated structures than L2 (disregarding differences in acquisition stages). The final median in L1 stems from only one data point (there are only 20 texts in L1), which might explain the odd behavior of the vas_no_prep and the no_subj median in this window. L2 median curves also level off, but at lower

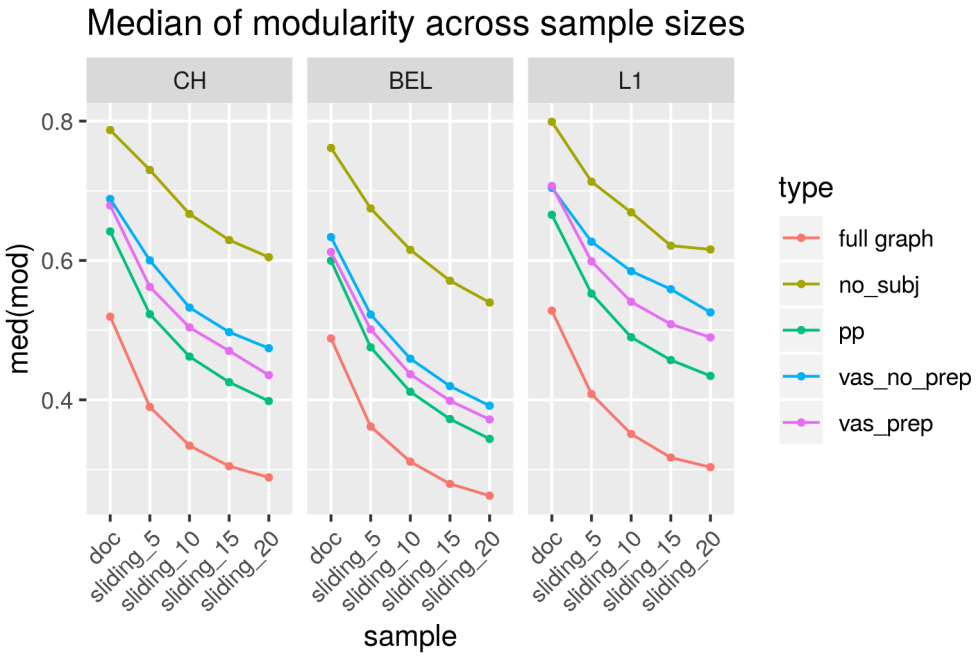


Figure 6.25.: Median modularity vs. sample size

levels than L1 median curves both begin and converge. Assuming that the median curves are steady and continuous, convergence in the whole BEL- and CH-population is likely to be reached at window sizes of between 30 and 50 texts, and a little sooner for l1.

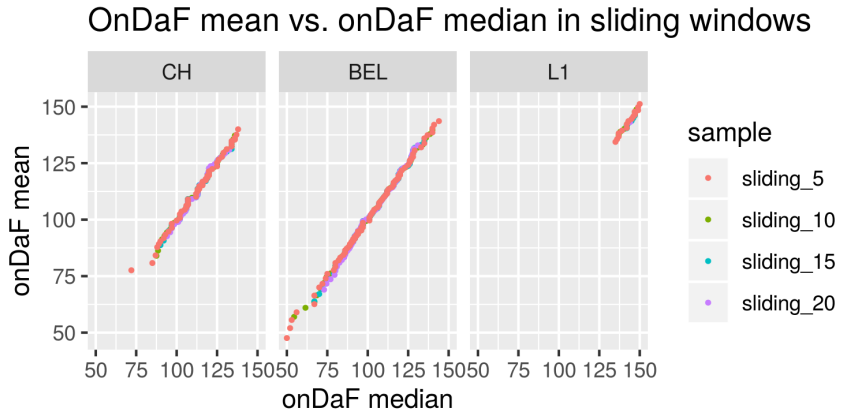


Figure 6.26.: onDaF median vs. onDaF mean in sliding windows

Fig. 6.27 shows that split by onDaF groups, convergence of the median is still not reached as early in L2 as in L1. This suggests that variance is overall higher in learners, as was predicted. The plot also shows that clear patterns, like the overlap of BEL-95 and BEL-130, emerge at sample sizes of 10 and go unchanged for larger samples. The overlap of the BEL-95 and the BEL-130 groups, and the lower position of BEL-115 is condensed evidence of the u-shape throughout corpus sizes. Unlike this, in CH, curves appear in ascending onDaF order.

Interestingly, in no_subj, L2 patterns do not stabilize like they do in vas_no_prep (fig.

6.28). Rather, medians cross at between 5 and 10 texts and again at between 15 and 20 texts. This may partially be due to the smaller size of a 20-text-window graph in no_subj. At 20 texts, it looks like the order of curves might be re-established in BEL (BEL-95 drops below 130 and 160 in the 20-text-windows). It might also be an effect of actual differences in the behavior of the measure in the no_subj graph. This cannot be validated internally, but requires fresh data.

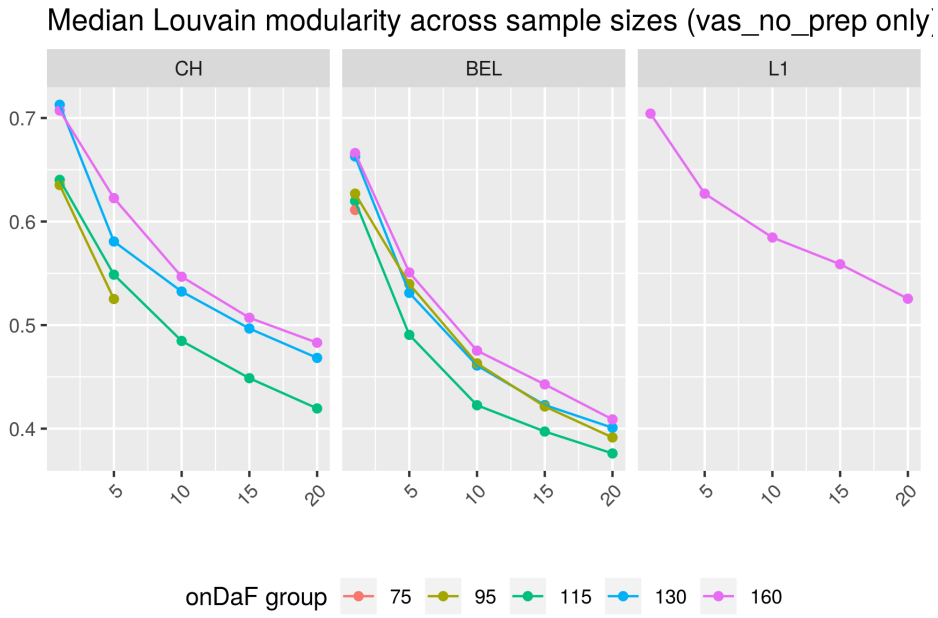


Figure 6.27.: Modularity median in corpus sizes by onDaF groups and language, vas_no_prep only. The final data point in L1 stems from only one window (there are only 20 L1 texts in Kobalt). Groups here are assigned by onDaF median of the window or by median of all documents in the corresponding onDaF group, which is why there are data points missing: There is no 10-text-window with a median onDaF $\geq$ 95.

Fig. 6.29 shows modularity values for the four window sizes and two specificities by language. Modularity is plotted against the onDaF median in each window, which is roughly equal to plotting against mean in this case since onDaF is distributed symmetrically across windows with exception of the very first data points in learner groups (fig. 6.26). Corpus size effects are clearly pronounced in that larger windows show lower modularity values, but the shapes of the distribution are similar for windows of 10 texts and higher. Fig. 6.30 is identical to 6.29, but includes regression lines based on an automatically fitted general additive model (geom_smooth gam on mgcv in R, Wickham (2016), Wood et al. (2016). Please note that unlike in the individual text analysis (which can also be considered an analysis of a sliding window of one text), these regressions are not statistically valid and are printed here only for a better illustration of an approximated trajectory, especially in the smaller sized windows, where a line diagram connecting each data point overemphasizes high deflections between data points. One of the prerequisites for the statistical validity of a regression model is the independence of data points, which is not given here because each sliding window overlaps with its neighbor in all but two texts (the leftward neighbor of the lowest ranking text is exchanged with the highest ranking text in the current win-

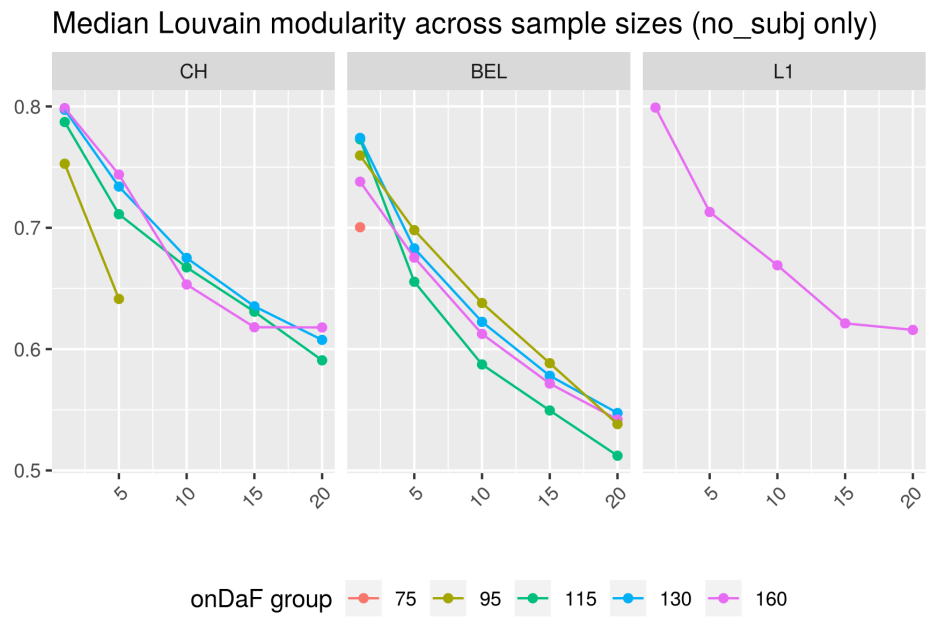


Figure 6.28.: Modularity median in corpus sizes by onDaF groups and language, no_subj only. Unlike in vas_no_prep, patterns between onDaF groups do not stabilize with larger corpus size in L2.

dow). This is why a sliding window analysis also implicitly gives some insight on the role of individual variance, not based on one, but two texts. Consequentially, data points in 5-text-windows are more volatile, because with each shift, 40% of the corpus is exchanged on average, while data points in 20-text-windows especially in the larger vas_no_prep graphs are very neatly situated on a clear trajectory that seems more informed by onDaF median than the 10% or 13% text exchange. This is relevant because it shows that individual variance, while it exists and plays a role in smaller windows especially, is not randomly distributed noise, but expresses a developmental trajectory.

Figs. 6.31, 6.32, 6.33 and 6.34 provide a side-by-side comparison of all windows by languages. To point out some observations:

- No_subj and vas_no_prep differ slightly in trajectories in BEL, and more strongly in CH. This is consistent with previous findings in this chapter; The no_subj trajectory in the CH-group is also more volatile and sensitive to changes in corpus size or grouping;
- Confirming the u-shape hypothesis in BEL, all sliding windows show the same drop in modularity values for intermediate onDaF score ranges in BEL-learners. Modularity begins to decrease at a window median of 75-80 onDaF points and reaches its lowest point at a window median of shortly under 100 onDaF points;
- For the CH-learners an approximation of a u-curve is visible for the 10- and 15-text windows, suggesting that if a u-shape exists, the data onset is at its lowest point. This point is reached by BEL-learners at around 90 onDaF-points, an onDaF range that is lacking of data in the CH-subcorpus. It might be that strongly decreasing

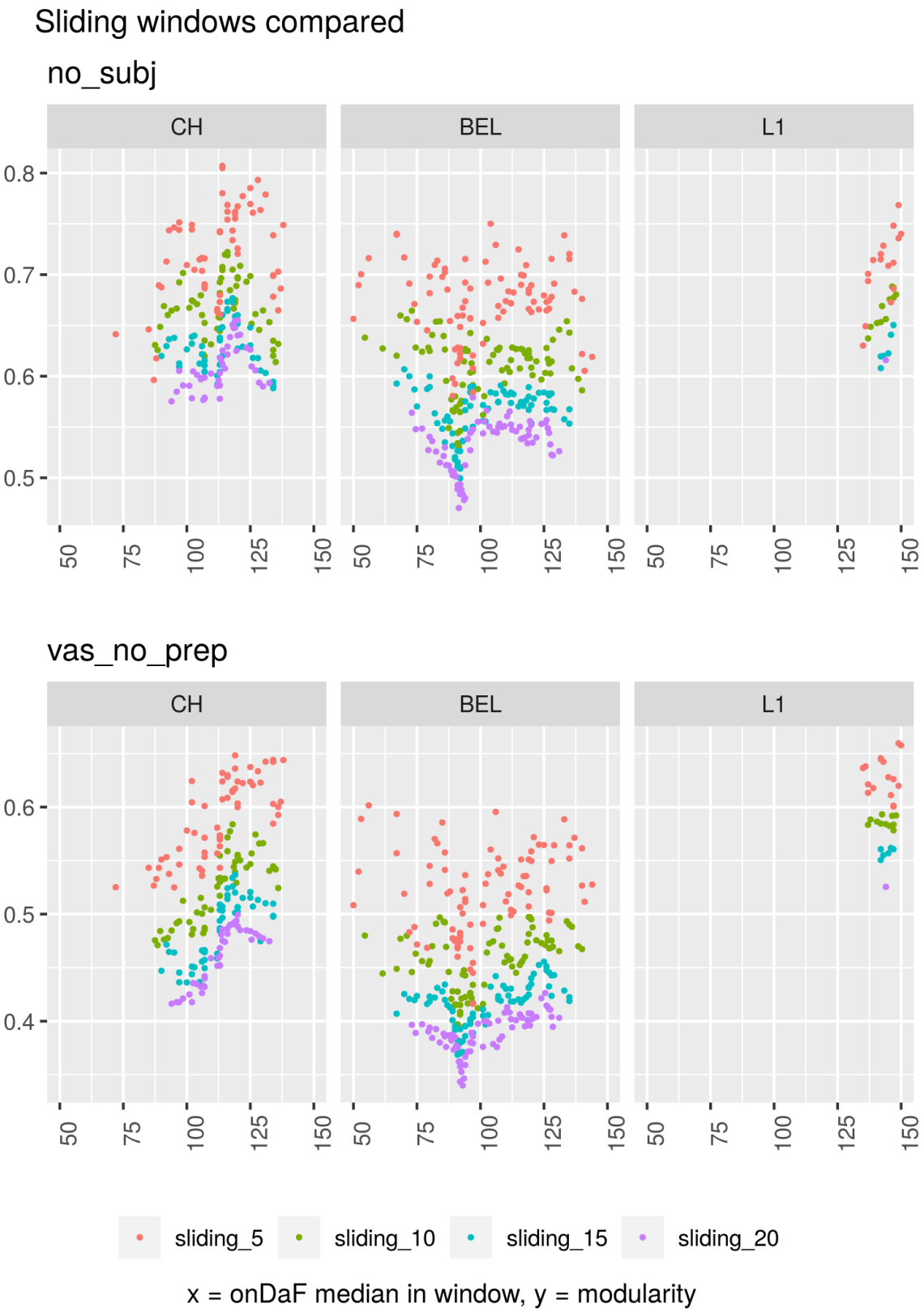


Figure 6.29.: Sliding windows compared

modularity effects are levelled in 20-text windows for both the BEL- and the CH-learners because the drop happens only within a relatively small onDaF range that is not ideally captured in the data and relativized through texts neighboring in rank but sufficiently higher in onDaF scores to tip the window into a higher modularity range overall. Also, for the CH-corpus, the 20-text windows begin at an onDaF median of around 90, which might be too late of an onset to capture the effect.

- However, judging from the combination of the (non-grouped) analysis of individual texts, the initial onDaF groupings, and the sliding window analysis here, it appears that possibly the CH learners behave differently altogether. While there are hints at a u-shaped learning curve in some of the analyses, those might represent a different phenomenon, such as an expression of greater individual variance through disparate writing styles, skill sets, or strategies, rather than a more or less grouped development, or can be attributed to a factor that is not considered in the analysis here. It is also possible that grouped vs. individual writing works differently in CH vs. BEL, and that different corpus sizes reflect different layers of emergence effects or lack thereof. Testing this requires more detailed modeling of expressions of emergence effects and idiosyncrasy in verb argument structures, and fresh data since this dataset is already overtested.
- For the larger windows and the upper scores, modularity values drop again after reaching higher levels with a maximum at around 120-125 onDaF points, for both the BEL and the CH windows. I will discuss this M-shape in relation to text length effects further below;

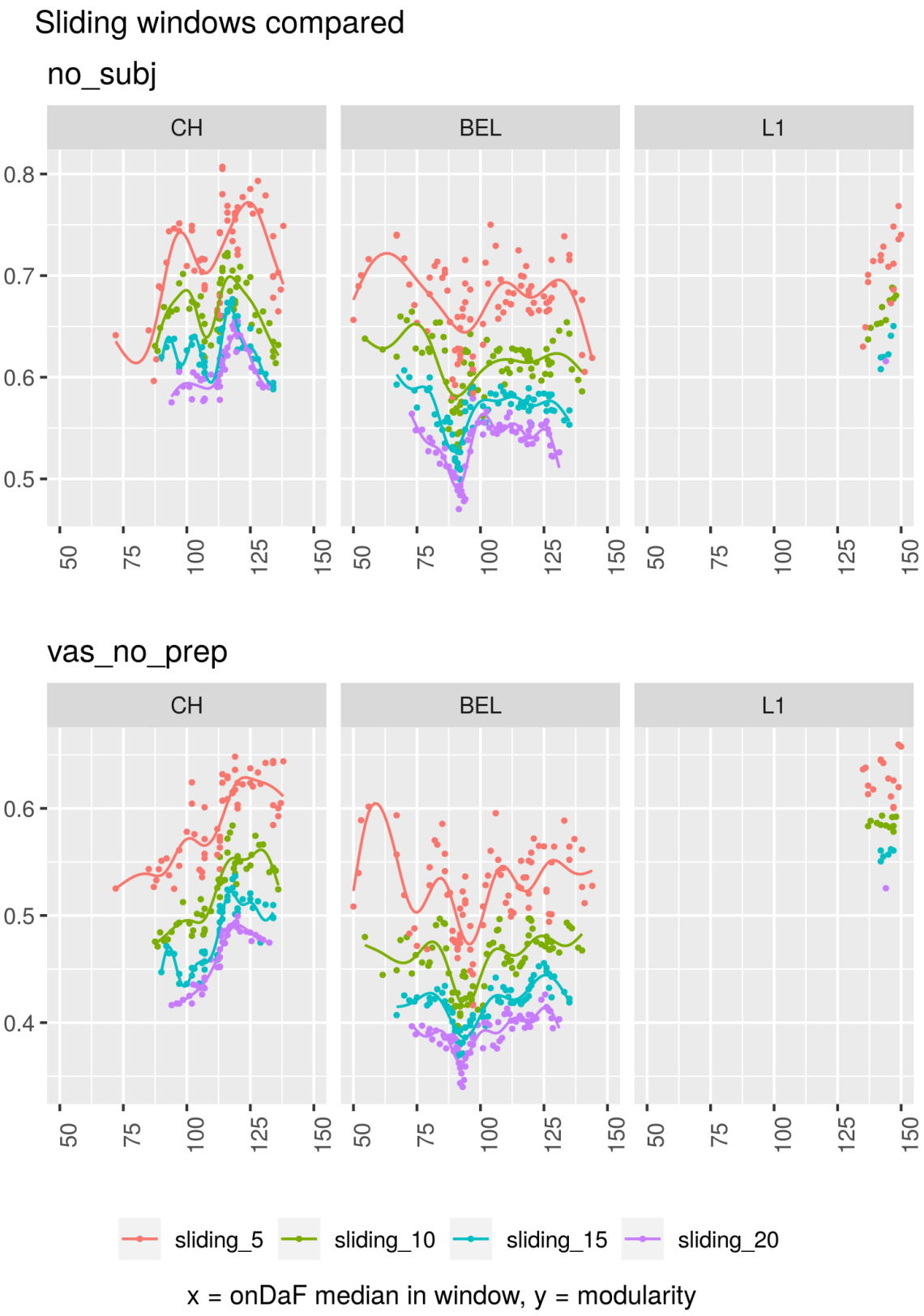


Figure 6.30.: Sliding windows compared with automatically fitted regression line. The regression is not statistically valid and plotted only for better discernability of trajectories).

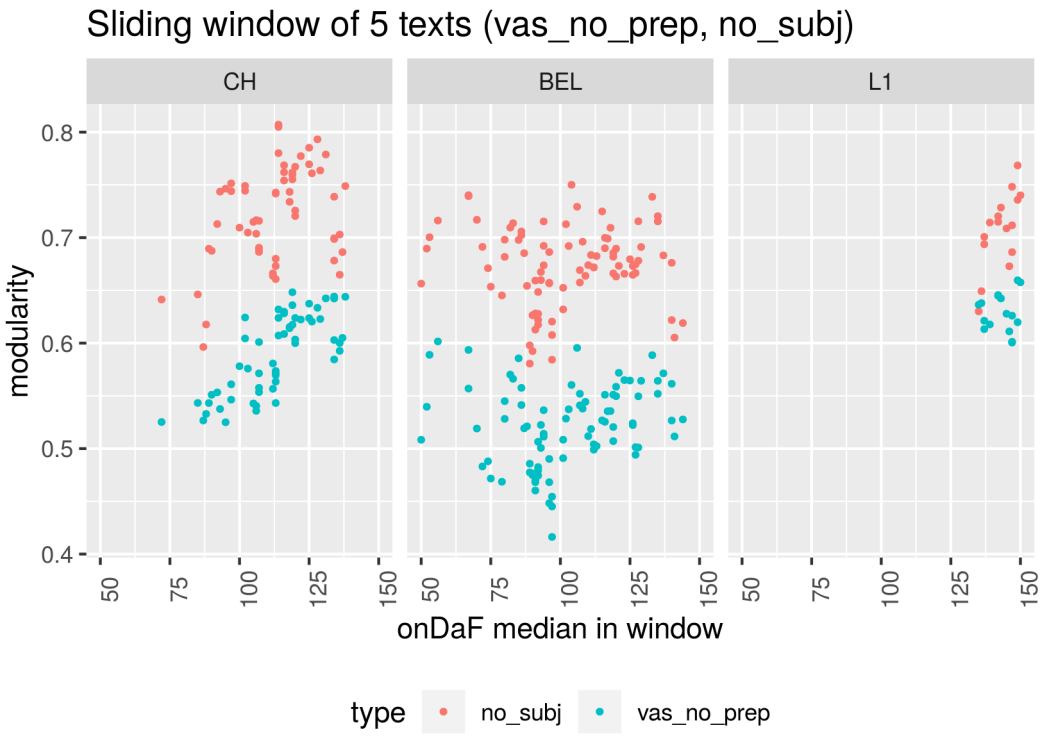


Figure 6.31.: Sliding windows of 5 texts compared

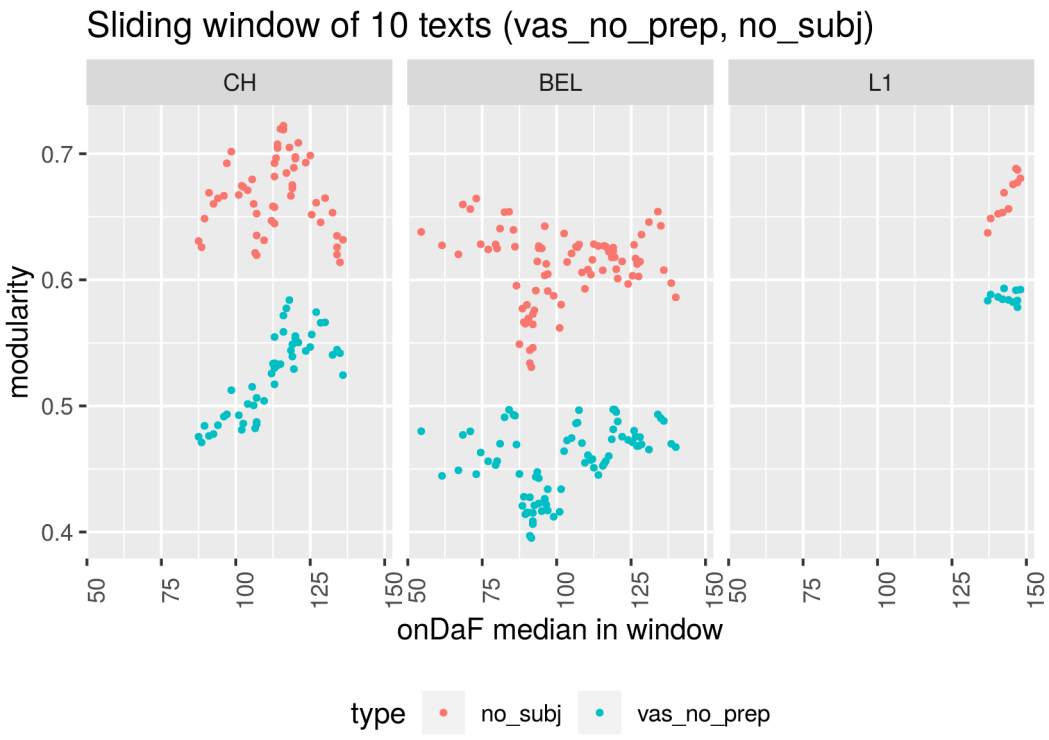


Figure 6.32.: Sliding windows of 10 texts compared

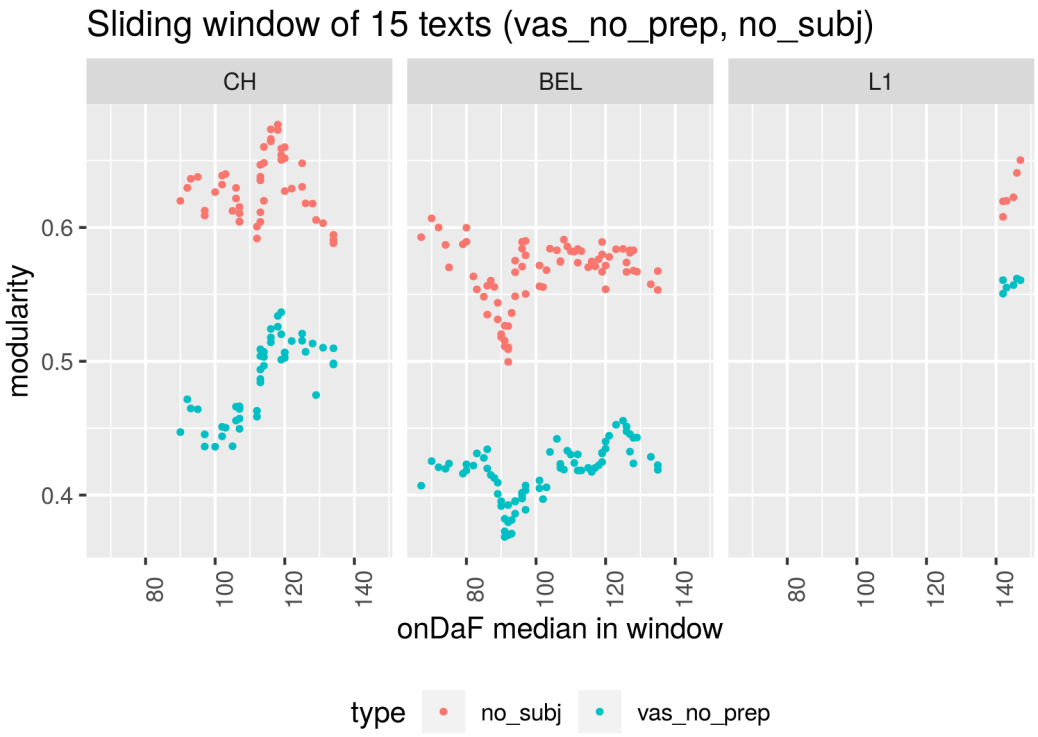


Figure 6.33.: Sliding windows of 15 texts compared

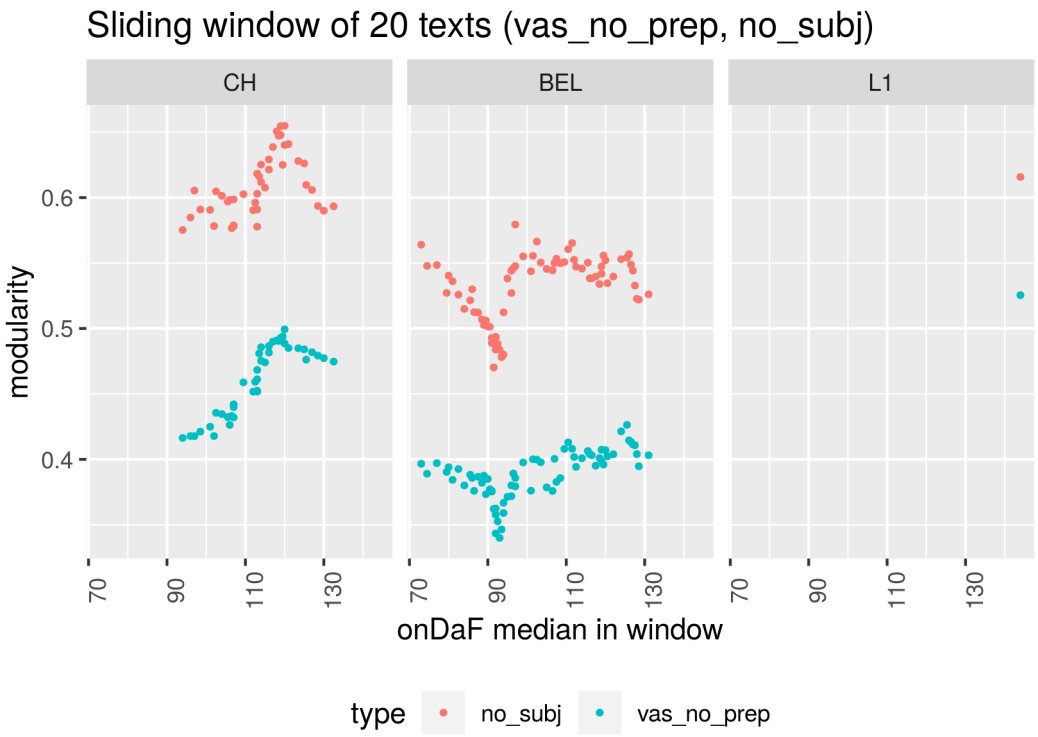


Figure 6.34.: Sliding windows of 20 texts compared

Looking through the sliding windows separately for each language, the progression to larger windows leads to more clearly defined and more similar trajectories between no_subj and vas_no_prep in both L2 groups (fig. 6.35, 6.36). Interestingly, in the larger windows, modularity values in vas_no_prep in both groups are higher at the upper end of the onDaF scale than the lower, which is not consistently the case for smaller windows. This suggests that *corpus modularity* is higher than *individual modularity* in these learners, i.e. that they are more similar in coselectional constraints.

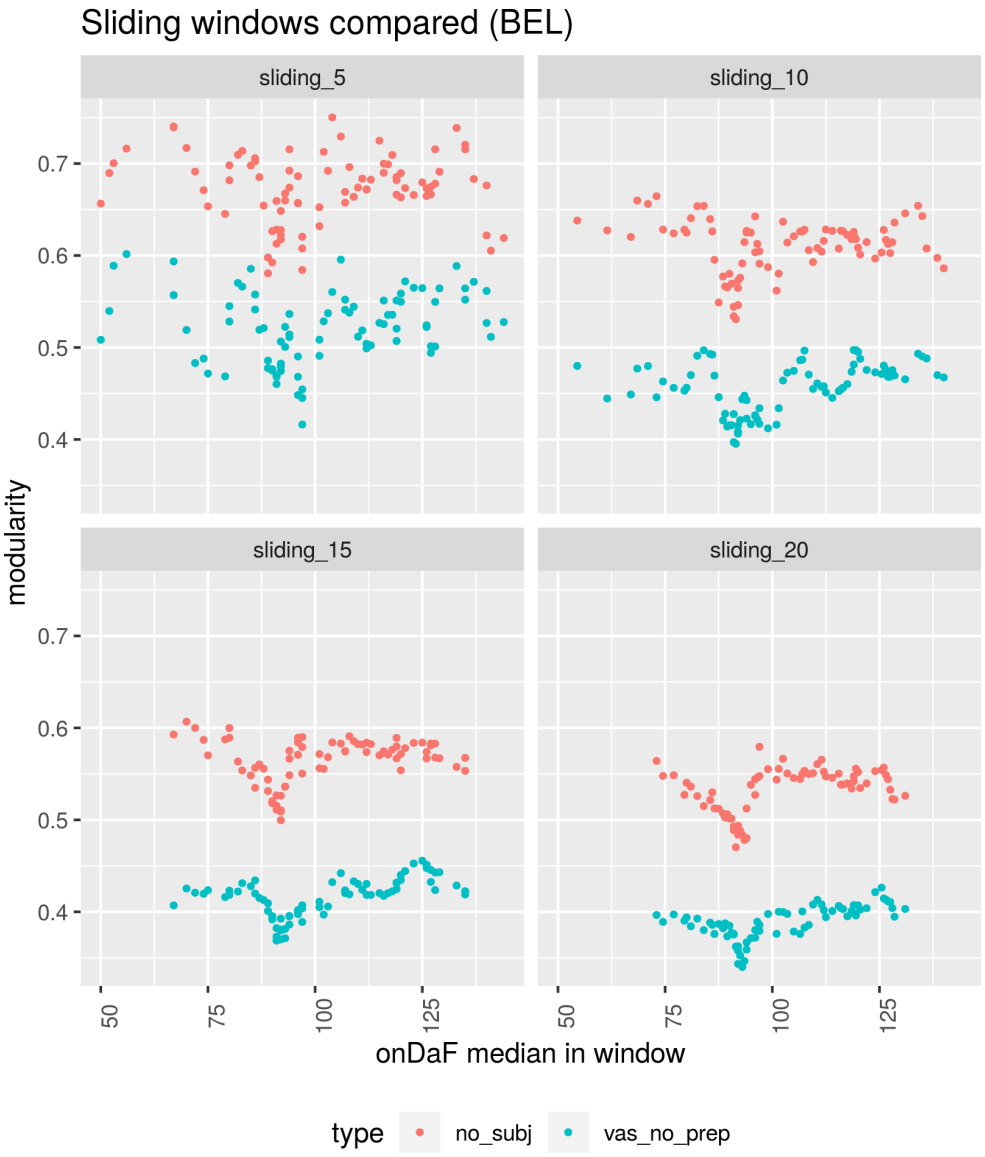


Figure 6.35.: Sliding windows compared (BEL)

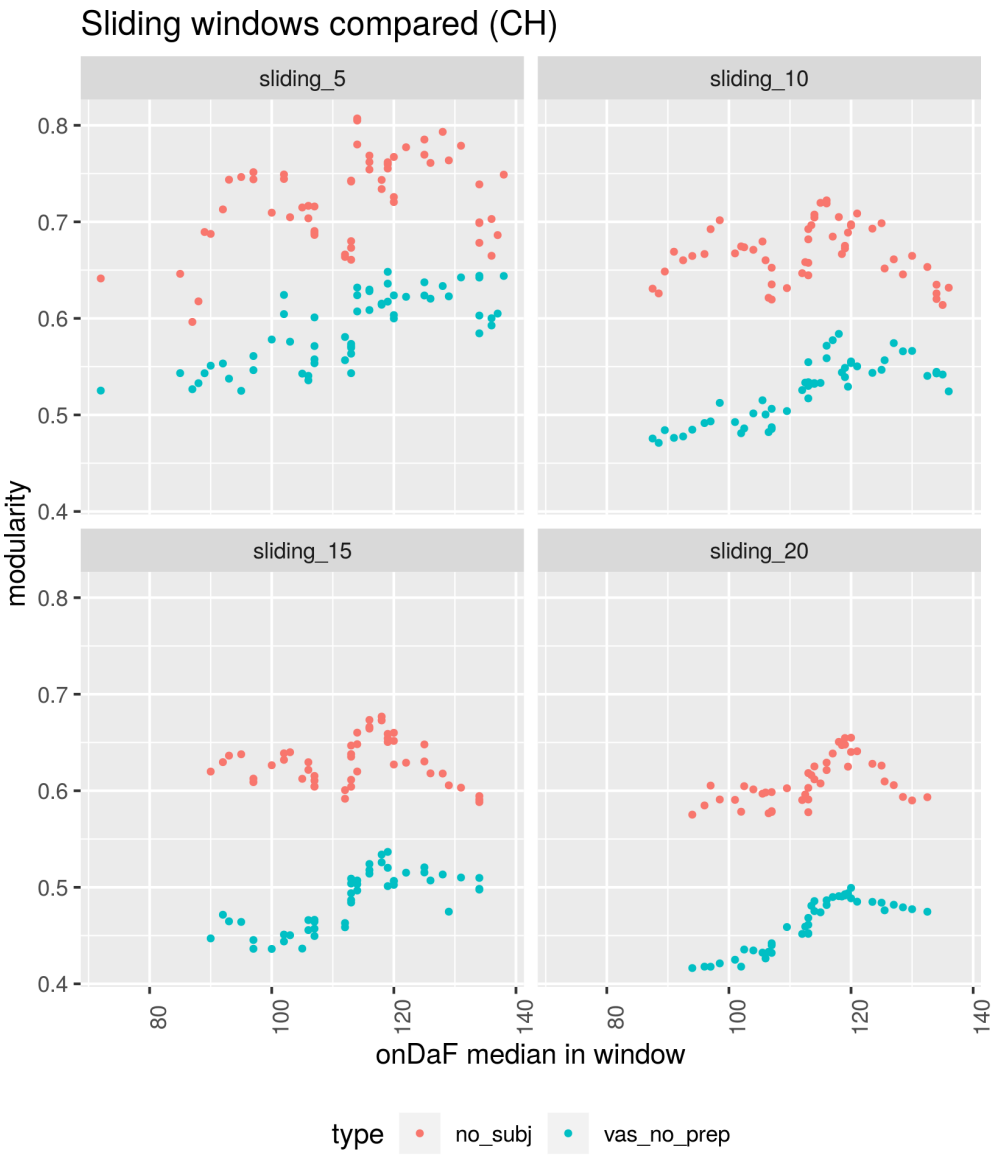


Figure 6.36.: Sliding windows compared (CH)

In L1 (fig. 6.37), modularity is more evenly distributed across onDaF ranges for the vas_no_prep graph, but there is a surprisingly strong effect in the no_subj graph, where modularity in 5-, 10-, and to a lesser degree also 15-text-windows clearly increases with onDaF. This is partially explained through corpus size (see section 6.3.4 for a discussion of text length), because L1-writers tend to write shorter texts with increasing onDaF scores as can be seen in fig. 6.38, where the higher onDaF windows group in the right bottom corner at lower text lengths, and the 20-text-window data point looks like it was forced to the meeting point of the two groups (top left vs. bottom right). However, if corpus size was the only cause for the effect, a similar trajectory would be expected for the vas_no_prep graph which is not the case (see fig. 6.37, sliding_10 and sliding_15, where modularity in no_subj grows, but stagnates or even drops slightly in vas_no_prep within the same

sample (window).[16]

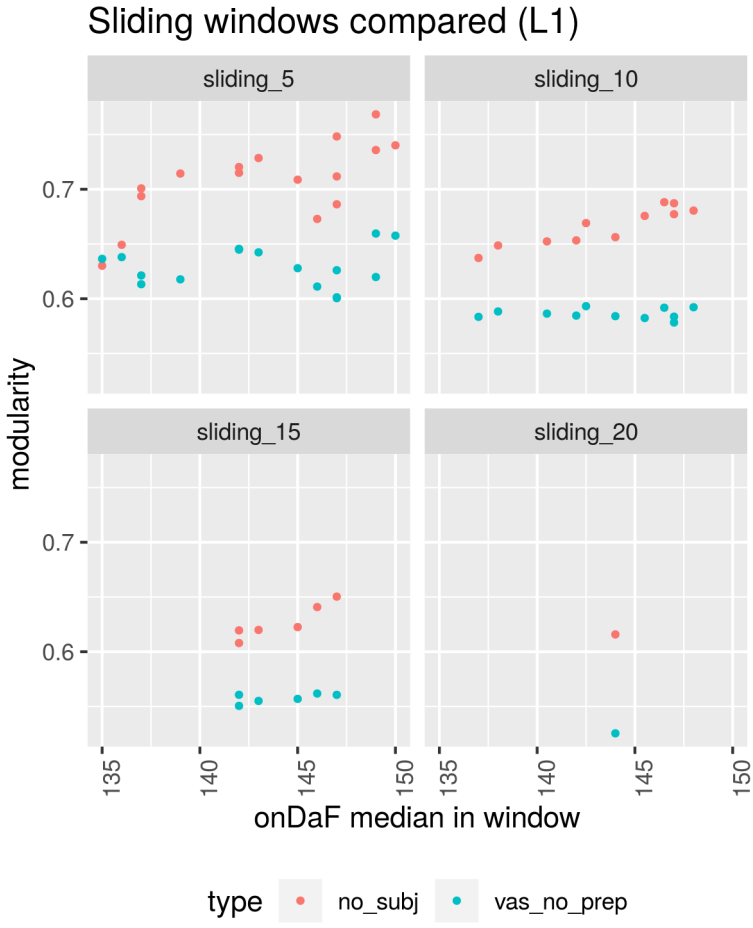


Figure 6.37.: Sliding windows compared (L1)

[16]Since there are only 20 L1 texts in the corpus, the 20-text window in L1 shows only one data point. This illustrates one downside aspect of sliding window sampling, namely that texts of ranks $n_{windowsize}$ to $m_{corpussize} - n_{windowsize}$) are included $n_{windowsize}$ times in the analysis, while the first and the last text are included only once, the second and the second-to-last twice, and so on. Therefore a sliding window analysis allows for a much more fine-grained analysis at intermediate stages, but not at the edges of the distribution. Such, in a 10-text-window, the final window is the same as the BEL-160 group in the initial onDaF-based group analysis.

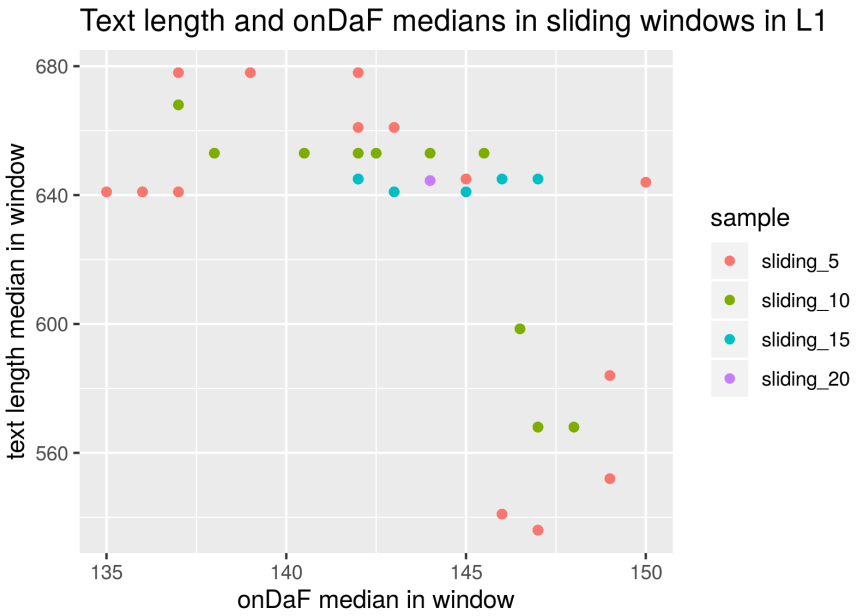


Figure 6.38.: Text length and onDaF medians in windows (L1)

Overall, the sliding window analysis cofirms the main findings from the onDaF-based analysis in section 6.1, in the following ways:

- u-shaped (v-shaped) curves are consistently observable in the BEL-corpus for windows $\geq$ 10 texts;
- modularity increases towards higher onDaF scores and in CH even surpasses L1-values in some graphs;
- modularity in the no_subj graph in CH shows large variance and perhaps two diverging tendencies around the same onDaF range as in the initial analysis (CH-130, 115-129 onDaF points).

With this, the ranking of texts based on onDaF neighborhood yields surprisingly clear trajectories for windows of 10, 15, and 20 texts in vas_no_prep and for no_subj in 15- and 20-text-windows, which corresponds, if not in perfect proportion, to size difference between the two corpora. It thus appears that onDaF-ranking is indeed linguistically meaningful, particularly where larger windows are compared, in that the variance in modularity between learner subcorpora is directional with onDaF. It appears then that a sliding-window-sampling functions as a simulation of an implied continuity in the data. At the same time, it confirms the results from the intital, grouped analysis, suggesting that a grouping by onDaF, both in a discrete and in a continuous design, yields consistent results and is not inferior to a continuous analysis. This is despite the shortcomings of dividing a continuous variable into discrete classes, and the doubtful status of the onDaF as valid language assessment in the context of lexicosyntactic development.

6.3.3. Summary: Corpus size and grouping

In this section, it was shown that

- non-random results in line with the hypotheses are observable for groupings of 10, 15, and 20 texts and in regressions over individual texts. Corpus sizes of 5 and 6 texts seem least productive and, unless based on a very dense (sliding window-based) analysis, seem to confound effects that are clearly observable in the other groupings. Perhaps this is an effect from interference of beginning emergence of grouped effects vs. too strong impact of individual variance.
- A sliding window analysis based on the ranks of the onDaF distribution yields clear and consistent trajectories. 5-, 10- and 15-text-windows are more similar to each other in CH than to the 20-text window, but not in BEL, where windows of 10, 15 and 20 texts are all more similar to each other than to the 5-text-window. It is possible that this is due to the 20-text-window in CH bridging the gap that exists in the lower-intermediate onDaF score range. It is also possible that 20 texts do span an onDaF range for many ranks that is too large for some parts of the scale and obfuscates combined effects from intra-group variance and transition points between acquisition stages.
- Overall, the results from the analysis of 10-text-samples based on onDaF-range groups were confirmed in the sliding window analysis and not challenged in the individual document analysis. This suggests that a grouping based on a c-test like the onDaF is sufficiently correlated with the aspects of learner language relevant to this study, and the trajectories in the sliding window analysis show that higher onDaF median in a window raises or lowers modularity systematically. The exact validity of the chosen cut-off points of the grouping cannot be determined in this study, but the groups do not seem to contradict results from more fine-grained groupings like the sliding windows.
- In several respects, it appears that individual vs. group effects may not be artifacts from insufficient control over the quantitative analysis, but expressions of different layers of emerging structures. This is particularly relevant to the future study of coselectional constraint in different corpora, but first requires a more detailed model of the dynamic processes involved.

To answer the questions from the beginning of this section:[17]

1. whether results from the onDaF-based grouping can be confirmed in other groupings, and whether coselectional constraint is detectable in individual documents:
    - Increasing modularity with higher specificity and similar trajectories for verb-specific graphs, but not no_subj, u-shaped trajectories in BEL and higher modularity values for L1 than L2 for most, but not the higher CH-groups are confirmed in sliding window samples and individual document analysis. A final drop in modularity in both BEL and CH after 125 onDaF points is observable. OnDaF10 is out of line with all other groupings and appears to be an unfotunate combination of grouping, low number of data points, and corpus size.

[17] All statements refer to the register and cohort reported. External validation on new data, and internal validation within such a replication study, is required.

2. how corpus size and modularity interact:
    - Absolute modularity values drop consistently asymptotically with increasing corpus size and converge at corpus sizes of 20–30 texts in L1 and, extrapolating from the curve, at likely 30–50 texts in L2 if grouped by BEL and CH without consideration of onDaF groups. Modularity for graph specificities converges at different values, clearly defined in larger corpora, consequentially values may overlap in smaller corpora (as in the individual text analysis in section 6.3.1.1, where modularities for no_subj and vas_no prep partially overlap). Median modularity in L1 converges at smaller corpus sizes compared to L2, pointing towards a higher impact of grouped (streamlined) phenomena, and at higher values, providing evidence for structural differences in lexicosyntactic graphs between L1 and L2 on average. However, results relative to each other (specificities, trajectories, differences and similarities between L1, BEL, and CH) are robust and stable in corpus sizes ≥ 10 texts for vas_no_prep.

      While at the point of convergence, corpus size effects are minimized, it appears that the largest possible window is not necessarily the best for capturing the trajectories in an imbalanced dataset, since the onDaF range covered by a larger window offers too much counterbalance to effects of decreasing modularity within a small range represented by only few texts.

3. whether and how linguistic (onDaF-based) grouping and modularity interact:
    - It appears that a sliding window technique is most successful at balancing out a dataset like this for internal validation, despite its reliance on fluctuating onDaF-ranges, but that onDaF-based grouping with sufficiently large groups is indeed capable of capturing the general trends very well even for relatively small corpora of 10 texts. Whether or not this is related to the reported incidental correlation of those groups with CEFR-levels cannot be confirmed or disconfirmed in this dataset due to skewed distribution.

      Whether the onDaF-10 range is a better grouping per se cannot be decided due to the conflation with smaller corpus size in the distribution of the data. High variance in the onDaF10-grouping, as it was reported in the 5/6-sampling, suggests that perhaps a 10-point onDaF range does not capture similarity well enough because the standard error of the test might be higher than 10 points and random effects are overemphasized in that view. A grouping by larger onDaF-ranges seems to counterbalance those sufficiently.

      A sliding window analysis of 10 texts is about as good as a an onDaF-grouped analysis based on 10 texts, while larger sliding windows mostly add continuity without greater changes to the implied trajectory.

4. which corpus size is best suited to capture the effects, and if any (and if so, which) of the corpus sizes analyzed here (1, 5, 6, 10, 15, 20 texts) serve as lower and/or upper bounds for best results:
    - A corpus size of 10 texts is sufficient in the onDaF-based grouping, a corpus size of 15 or 20 texts in the sliding window analysis yields the clearest trajectories and suffices for convergence in L1 but not L2. It appears that in a dataset of roughly 80 texts for the BEL-learners, even collected modularity values from individual texts are in line with the trajectory, but for the grouped data, 10

texts seem like a reasonable lower bound. In the CH data, 15-text-windows seem to represent an upper bound at which effects in lower onDaF ranges are not swallowed by blending those texts with those of higher modularity in a common corpus, but this is likely idiosyncratic to the Kobalt dataset. It also appears that a corpus size > 1 and < 9 is not ideally suited due to interference of individual and emergent effects.

- Assuming convergence of modularity values after 20 or 50 texts in L1 vs. L2, the existence of an upper bound at which differences by onDaF or other factors become too unpronounced against corpus size effects cannot be conclusively verified or disconfirmed from the small dataset at hand. It is possible that differences submerge into smaller numerical ranges at smaller variance in the mechanics of the measure, but it is also possible that they are drowned out by hyperconnectivity related to lexicosyntactic productivity. In sufficiently large corpora, new edges between many arguments and verbs will be created, constantly shifting the balance between idiomaticity and productivity, or higher and lower modularity. From the data here, an upper bound is not observable, but the corpus is also quite small. It is possible that in a larger corpus, the difference between weighted and unweighted modularity plays out differently.

6.3.4. Text length

Some of the sliding window analyses and the initial grouped analysis show a drop in modularity after 120–125 onDaF points in both learner groups. This raises the question of text length or corpus size in tokens as a potential confounding factor. Obviously, longer texts, especially if they cluster within certain onDaF ranges as is the case for BEL, impact corpus size, and since corpus size interacts with modularity, an interaction is to be expected.

6.3.4.1. Text length as a construct

Text length is often discussed as a confounding variable in corpus linguistics, since, naturally, any quantification of word coselection will depend on the chance of (co-)occurrence of words. It will therefore interact with not only corpus size as measured in texts, but also the number of tokens included in those. Based on the long-tailed distribution of lexemes in natural language, the distribution only stabilizes after a certain text length (in a very short text, word frequency may even be equally distributed, like in this sentence until the last comma). A certain text length is thus required to fully unfold the distribution to a point where even the more common verb argument co-occurrences are included, and any cut in text length will result in losing most of the tail and therefore a relevant part of the interesting lexical material.

But there are still more complicating factors. Text length itself is a weak construct in several ways. First of all, measured in tokens, an increase in text length can be grounded in changes related to very different linguistic concepts, such as the number of analytical realizations of TAM, the number and kind of modifications, number of propositions, ideas or rhetorical structures, or simply the split writing of finite particle verbs in German. If two previously identical texts were increased by 20% of their tokens, it would be impossible to tell from this information alone how they had changed and whether they were still similar. One might argue that this is only true for individual texts, and that with sufficiently large corpora, distributions will emerge more clearly. However, this is likely only

true of canonical language (longer sentences in newspaper language may have specific implications), but not necessarily in learner language: In BEL, onDaF scores can be guessed reliably from text length, thus it may also be possible to predict linguistic correlates. But in CH, the same is not true, learners of all proficiency levels write texts of approximately the same length. Some may reach final length through coordinating verbs and arguments to long lists, while others may fill the space with more TAM constructions or modification.

Secondly, text length is not an independent variable. A text does not just happen to be longer than another. Rather, it is a function of genre, register, style, propositional density as in the number of arguments, stories, agents or characters included in the text, rhetorical structure (for example a list of activities may take fewer tokens than the same amount of information presented in a narrative structure), speakers' idiosyncrasies and preferences, and cognitive factors including stage of acquisition.

The amount of text that can be produced in a given setting is limited by at least three factors: How much do I *want* to write (inspiration, motivation), what *kind of text* am I writing (genre, register) and how much *can* I write given the context and my linguistic and cognitive means (degree of automation, writing experience, cognitive overload threshold, amount and complexity of text that can be kept and handled by working memory). It is also a matter of general proneness to writing activities, and more or less successful planning.

These factors all interact with writing experience, which is an interesting variable in this corpus: The learners are typically 3rd or 4th year university students, while the L1 writers are 12th grade high school students, meaning they are younger, less educated, and arguably at a lower level in their L1 acquisition and their cognitive development, certainly when it comes to argumentative and abstract reasoning and expression (see chapter 2.1.1 for a discussion of the role of late L1 acquisition in this context). They are also less experienced in writing 90 minute exams, and they do not belong to a self-selected group of people choosing to pursue university studies in a language subject. Perhaps they even represent a group of young people *less* attracted to writing, since the data was collected in a high school *Grundkurs* (3 weekly lessons, lower influence on final grade, final high school exams in subject facultative) as opposed to *Leistungskurs* (5 weekly lessons, higher influence on final grade, final exams in subject obligatory). All of these are overall confounding variables, not only with respect to text length. But since text length is presumably easy to quantify and compare, they are easily overlooked in this respect.

A longer text is also not necessarily an extended version of a shorter text, and does not have to stay the same kind of text from beginning to end, but can change through its course. This includes the incorporation of new insights or ideas during the process, but also changes in style, register, or even genre. This means that cutting a text after n tokens does not necessarily only shorten it, but may have repercussions on the representation of structural and stylistic aspects in the analysis.

In the BEL-corpus, it is not uncommon for speakers to begin their writing in a relatively argumentative and neutral tone and then shift to a much more passionate, emotional and even politically and/or morally charged tone mid-text, as is illustrated in examples (a) and (b).

(a) "Meiner Aussicht nach ist das eine sehr strittige Frage, alles hängt davon aus, von welchem Standpunkt wir sprechen wollen. Wenn wir materielle Sachen in Betracht ziehen, dann seht das Leben heute bestimmt ganz anders als früher."

'In my view, that is a controversial question, it all depends on which perspective

we want to speak from. If we consider material things, then life today looks very different from earlier times' (BEL_020, sentences 14 and 15, rough translation with ZH2-like adjustment of lexical choices that are semantically not L1-like).

(b) "Es war schwer, aber Freundschaft, eine positive Einstellung, die Hilfsbereitschaft der anderen halfen immer! Die Wirklichkeit ist so wie sie ist! Wir müssen menschlicher sein und nicht nur so scheinen!"

'It was hard, but friendship, a positive outlook, other people's readiness to help were always helpful! Reality is what it is! We have to be more human and not only appear like it!' (BEL_020, sentences 40–42, rough translation)

Although both (a) and (b) are quite vague contentwise, (a) clearly adheres to a more argumentative register lexically and syntactically (*Meiner Aussicht [Ansicht] nach*, 'from my perspective', *strittige Frage* 'controversial question', *in Betracht ziehen* 'consider', syntactic embedding, which is not necessarily argumentative, but more formal and conceptually written; explication in subordinate clauses), while (b) belongs to a much more conceptually oral register[18] reminiscent of a motivational speech or a sermon, which is expressed in shorter sentences, coordination rather than embeddings (both within clauses in the list 'friendship, a positive outlook, other people's readiness to help' and between sentences). Speaking anecdotally from my observations in the data, shifts like these are unidirectional both in that a register change once manifested is unlikely to shift back to the earlier register, and that I am not aware of a text in the corpus that starts out highly emotional and ends in an argumentative and sober tone. This is not to say that other shifts are not possible and do not occur even in this data though, these are perhaps just the most salient ones. In a longer text, normalizing for text length then can also imply unintentionally normalizing for register for only the longer texts (because shifts in shorter texts are more likely to occur before the token limit than they are in longer texts).

The words chosen, their order of appearance, and their frequency are not random either, but depend on the selection of (rhetorical) arguments and topics, and many stylistic choices. Some of those choices may be subconscious, but others are part of the planning process: A learner knows how much time they have to complete the task, at least some of the ideas they would like to incorporate in their writing (some observable meandering of thoughts aside), and apply their knowledge and experience with writing, including genre and style, to the composition of the text. In a text that is written to be concise, a lot of new information (i.e. different lexemes) will be presented within a small token range, while in a text that is designed to be longer, some ideas might be dwelled on for a longer time, lowering the novelty rate or type-token ratio.

The question of what coselection of verbs and arguments looks like for an equal amount of text therefore requires an operationalization of the concept of a specific amount of text to ensure comparability, which is not trivial. A text is not simply a collection of characters, words, or sentences, but a somewhat holistic object. Sampling parts of a text and treating it as if it was the same as a shorter text does not do justice to this fact. This is easy to see

[18] The distinction between the conceptual and medial realization of a text in the German linguistic discourse goes back to Koch and Oesterreicher (1985), who theorize that any expression of language can be located on two independent continua, one referring to its expression of personal closeness vs. distance the other referring to its realization in written vs. oral forms. This was to counter the frequent misconception that written speech is necessarily more formal. Rather, a formal register may also be realized in spoken form (as in a parliamentary speech) and an informal register may also be realized in written or printed form (such as a messenger text).

on a sentence level: The two sentences in the following example are both second sentences in their respective texts and declare similar points, differing, aside from the repetition of the prompt in (2), mostly in the scope of the authors' statements, where (2) references a discourse that presumably exists outside of the essay, while (1) rates the question itself as difficult and restricts the perspective to the authors answer. If both sentences were cut after 14 words (disregarding the quote tokens in (2)), the semantic similarity would completely disappear, since (2) would be cut after 'glaube ich' (I believe).

(1) Das ist wirklich eine schwere Frage und eindeutig kann ich auf sie nicht
that is really a difficult question and definitively can I on she not
antworten. (BY_070)
answer
'That is really a difficult question, and I cannot give a definitive answer to it'

(2) Zu dem Thema "Geht es der Jugend besser als der früheren Generation"glaube
To the topic "Goes it the youth better than the earlier generationthink
ich, eigentlich gibt es keine Lösung, weil andere Leute an die andere Seite
I, actually gives it no solution, because other people on the other side
denken. (CH_034)
think
'I believe there is actually no solution to the question "Does the youth do better than the previous generation?", because other [different] people think about the other side [different aspects]'

Examples (3)–(5) illustrate the same problem for lexical material, where the verb-argument coselection of *entstehen* and *Problem* ('emerge, arise', 'problem') reappears identically, but at different points in the argument: Once at the beginning ((3), sentence no. 4/verb at token 39), once mid-text ((4), sentence no. 25/verb at token 178), and once in the last third of the text ((5), sentence no. 37/verb at token 514).[19]

(3) Inzwischen entstehen viele Probleme mit der heutigen Generation, oder
Meanwhile emerge many problems with the today's generation, or
sozusagen der Jugend heute. (CH_055, sentence 4)
so-to-speak the youth today
'Meanwhile, many problems emerge with today's generation, or so to speak, the youth today'

(4) So wie neue Probleme entstehen und entdeckt werden, werden andere
Such like new problems emerge and discover.pass be.aux, be.aux others
auch bekämpft und angegangen. (DEU_018, sentence 25)
also fight.pass and approach.pass

[19] Example (5) is taken from the original Kobalt project corpus (Zinsmeister et al., 2012). All ZH1 annotations as they were encoded by the Kobalt project were left unchanged, where in this case the grammaticality is questionable and can only be maintained if 'wir leben' was viewed as a semantically somewhat unrelated parenthetical insertion ('In der Zeit entsteht ein Problem der Einsamkeit, und wir leben', 'In that time a problem of loneliness emerges, and we live') instead of an intended relative clause ('in der wir leben', 'that we live in'). For the purposes of this study, cases of uncertain grammaticality such as this one are not of greater importance, since only structural, hence no lexical, material would have been added on a ZH1 annotation level anyway, and the lexemes of functional word classes such as demonstrative or relative pronouns reliably appear with sufficient frequency to be noted, even if a few cases were missed.

'In the same way that new problems emerge and are discovered, others are being approached and combatted'

(5) In der Zeit, wir leben, entsteht wirklich ein riesiges und globales Problem der
In the time, we live, emerges really a huge and global problem of
Einsamkeit. (BEL_013, sentence 37)
loneliness
'A huge and global problem of loneliness really emerges in the time that we live in'

The combination of these two lexemes is common and arguably lexicalized. A search in the German reference corpus DeReKo (2019-I, Kupietz et al. (2018)) returns roughly 1900 results for the exact phrase 'Probleme entstehen' not accounting for any of its morphological or syntactic variations. It occurs seven times in the Kobalt corpus, once very early in the text, four times in the medium section and twice relatively late (sentence no. 32, BY_082, sentence no. 37, BEL_013). Cutting the texts at a token threshold is therefore not ideal, since it means losing lexicosyntactic similarities that exist even relatively frequently in the writing simply because they appear later, like cutting a ying yang symbol in half and concluding that it was mostly white. This could be viewed as a question of required sample size to reach lexical representativity. But I would argue that, given the Zipf-distribution of lexemes, the non-randomness and typical length of argumentative text, it is unrealistic to reach lexical or lexicosyntactic representativity on a topic in a single text of any length.

In that sense, if an equal amount of text refers to a unit of how much text is produced within a certain time frame and in response to a certain prompt, then it has been accounted for in considering the full texts for the corpus, where text length is given in units of text, namely one per writer.[20]

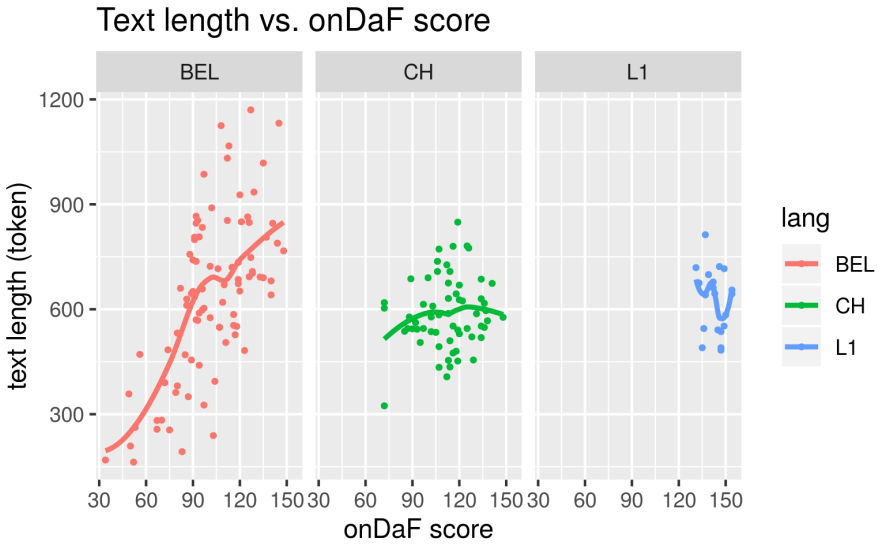


Figure 6.39.: Text length distribution by onDaF-scores

That said, not adjusting for text length in the data here leaves room for doubts about the validity of modularity trajectories in the BEL-corpus and its comparability with the other corpora. In BEL-learners, text length strongly correlates with onDaF scores in that

[20] More would be unlikely in this scenario, but not entirely impossible, for example if they first wrote an argumentative text and then followed up with a story to examplify their argument.

more advanced learners write increasingly longer texts, the longest text reaching including seven times as many tokens as the shortest (factor 7.17, 1170/163 tokens, see fig. 6.39). The same is not true of the CH-group, where the longest text is less than three times as long as the shortest (factor 2.62, 849/324), and the native speakers write texts that are all longer than most intermediate learners but much shorter than the advanced BEL-learners, and with even less variation in text length than the CH-group (factor 1.68, 813/483). This means that in the usage of an early intermediate BEL-learner, combinatorial power is also restricted by the low number of tokens in their usage.

Does this reflect overall or potential constraint in verb-argument coselection? This cannot be answered definitively by a corpus study like this, because it is impossible to tell what could have been if learners had written longer texts. If an early intermediate BEL-learner absolutely had to write a text of 1200 tokens, would they repeat over and over what they had written before? Would they go more in depth? Would they list more things that can be 'had' or 'done'? Or would they come up with a set of new vocabulary for a text that is planned out longer? From the shorter texts alone, we cannot tell, and with a focus on *natural* language production in corpus linguistics, we arguably do not want data that is enforced in this way, either.[21]

The modularity values here are therefore best described as a lower bound for interconnectivity or an upper bound for constraint: The writing of the group represented in the corpus is *at most* as constrained in terms of lexicosyntactic cooccurrence as the modularity value suggests. Language is always situated and structured through the context of use, and no two contexts are exactly the same, meaning that creating exact comparability, although very much desirable for a quantitative study, is very difficult. To truly exclude effects of text length it would be necessary to collect data that asks of learners to write texts of a certain length and study the variation in genres, registers, syntactic and lexical material they produce; Or at least, in this data, to annotate and account for register, genre, rhetorical structure, modifications, syntactic complexity, and propositional density and then explain the interrelation of text length and modularity from there. This is unfortunately impossible within the scope of this study. However, since as can be seen in 6.40 an interaction between text length and modularity exists and text lengths are so strongly correlated with onDaF in BEL, an assessment of the impact of text length seems necessary to validate the previous results and to see if text length can explain the final drop in modularity that was observed in the previous analyses in most advanced learners, and in no_subj in particular.

Figs. 6.40 for both learner groups and 6.41 for L1 (please note the free y-scale between graph specificities) show that there is a relationship between text length median of the texts included in a sliding window and the modularity of the window. In BEL-learners, three groups can be observed in the plot: short texts with high modularity and low onDaF score, long texts with low modularity and intermediate onDaF score, and long text with high modularity and high onDa>F score. CH-learners have higher modularity values with higher onDaF scores at steady text lengths across the corpus. Interestingly, an onDaF effect for text length and modularity exists even for the L1-group (fig. 6.41), where at a text length median of 650 tokens in a 10-text-window (645 tokens in a 15-text-window), modularity varies between 0.63 and 0.65 (0.6 and 0.63 respectively) almost neatly arranged from lower to higher onDaF median in the window for the no_subj graph

[21] Of course this is not entirely true. Much of the learner corpus data, and in fact much of learner language, is quasi-experimental rather than natural. This applies to exams as much as controlled corpus data collection.

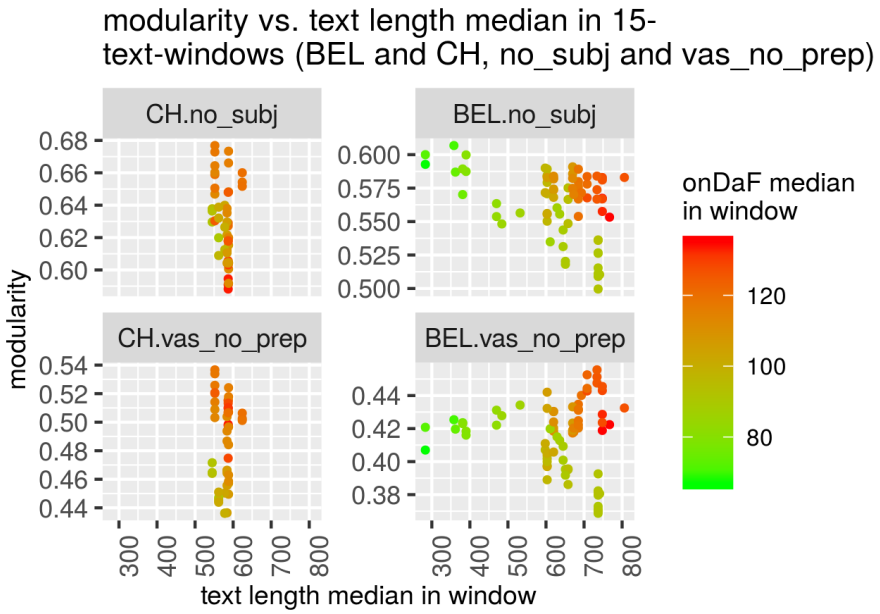


Figure 6.40.: Modularity vs. text length median in 15-text-windows (L2)

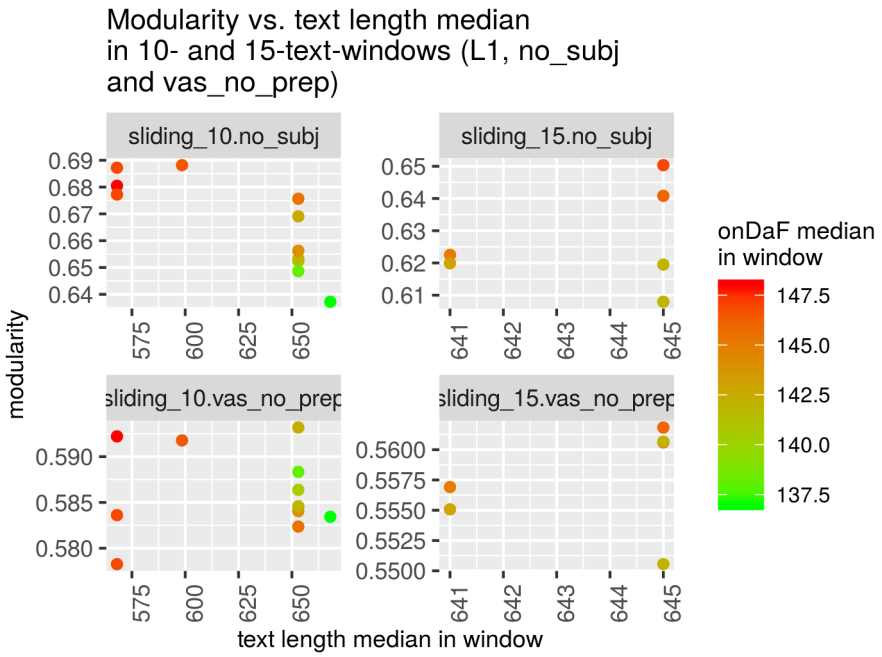


Figure 6.41.: Modularity vs. text length median in 10- and 15-text-windows (L1)

type. The vas_no_prep graph type does not follow the same pattern. However, since text lengths vary much less in L1 than L2 and are moreover devided into a short and a long text group (< 600 vs > 650 tokens), this result should not be overinterpreted.

6.3.4.2. VAS-based text length normalization

Since a token-based text length normalization is linguistically underspecified, it is not expected to provide much insight. Figs. 6.42 and 6.43 show that, indeed, the distribution and the trajectory barely differ between the full text and a version cut after 450+ tokens,[22] except for slightly higher modularity values in the no_subj graphs.

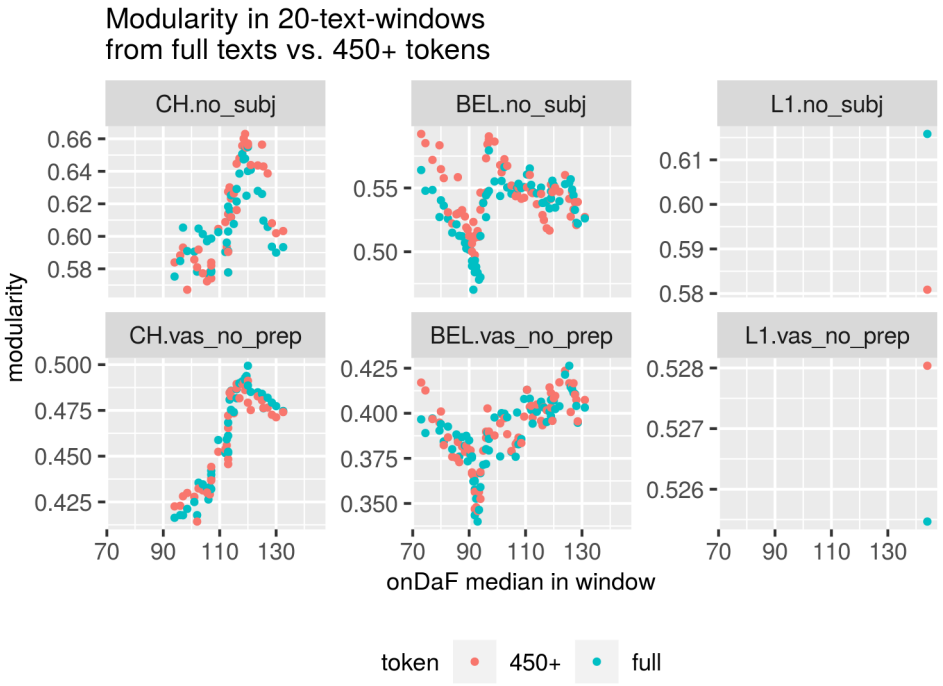


Figure 6.42.: Modularity in 20-text-windows based on 450+ tokens and full texts, free y-scale.

A normalization is thus performed based on VAS. For this, a fixed number of verbs (40) with all their arguments is considered in the analysis. This is close to a definition of propositional density or information unit. Sampling for VAS instead of tokens works as a noise reduction in text length control, because it frees the analysis from factors interacting with the text length that have little to do with the lexical constraints on verb argument structures such as the tendency towards higher syntactic complexity and more modification in more advanced learners (although it does not do so entirely if one considers nominal vs. verbal style as it is described for academic and formal registers (Biber and Gray, 2011; Petersen, 2014; Fang et al., 2006)). However, it is not fully unproblematic, either: If accounted for by verb, the more advanced texts will still have more lexical variation, argument constraints, and less repetition, i.e. higher modularity, through the fact alone that they include more complex vas, i.e. more edges. Limiting the number of argument slots instead has the reverse effect, where fewer verbs are represented in the analysis. Sampling only the most frequently used verbs is an option to compare their flexibility. But it does not necessarily reflect the overall modularity of VAS usage, and certainly does

[22] 450+ denotes a number between 430 and 480 where sentence boundaries were respected. The cut-off is arbitrary, but was designed to be inclusive in the sense that not too many texts should be much shorter for a valid groupwise comparison. Since the normalization does not seem to make much of a difference, further results will not be reported here.

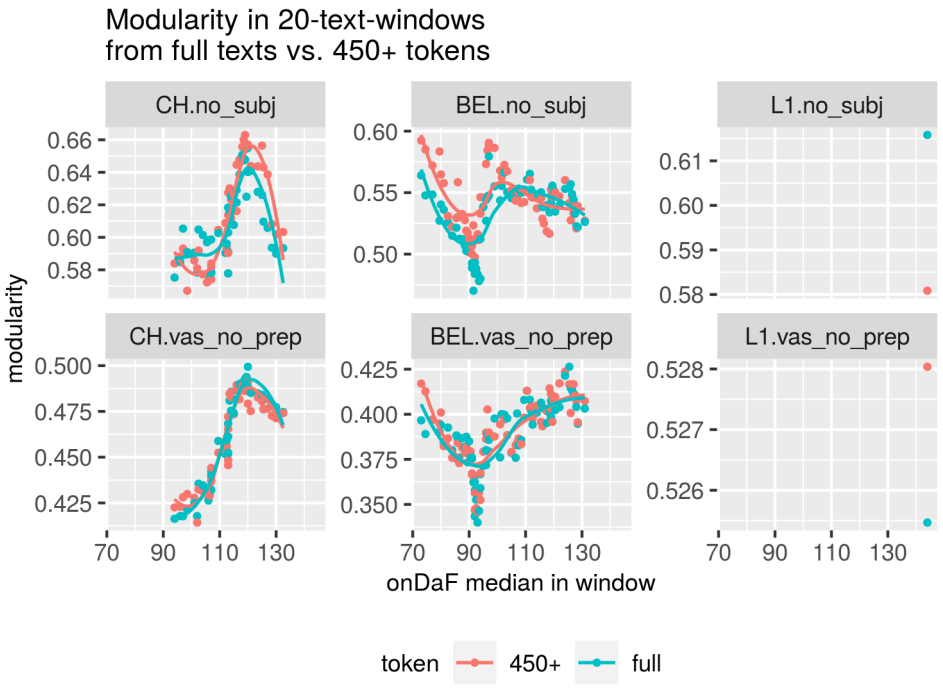


Figure 6.43.: Modularity in 20-text-windows based on 450+ tokens and full texts, free y-scale, with approximate trajectory

not reflect the differentiation process of vocabulary growth, because for a graph that has *both* the repetition of the common words *and* lexical differentiation, a limitation by VAS can only capture half of each or an unbalanced amount of both. However, looking at the development of constraints in the coselection of verbs as governing units, the most meaningful way of limiting text size seems by sampling for a fixed number of verbs and including all of their argument slots, even if those are changing patterns.

Interestingly, for the BEL-corpus, the number of verb argument structures differs between learners by an even larger factor than the token number (the maximum for argument structures in a single text being 169, the minimum 19, factor 8.9 vs. 7.18 for tokens), while for the L1 and CH-groups, they are much more closely distributed (CH: 96/37, factor =2.59 vs. 2.62 for tokens, L1: 93/57, factor 1.63 vs. 1.68 for tokens). While the variance is high between learners especially in the BEL-group with a range of 19–169 VAS, fig. Fig. 6.44 shows that a median is relatively stable across onDaF ranges and that for most texts, a number of 40 VAS represents an actual sample (and not the whole of the text, somewhat nivellating the structural effects previously discussed) while still including the BEL75-group. Again, 40 VAS does not constitute any kind of theoretically plausible unit, but is chosen pragmatically in the context of the data.

Figs. 6.45 and 6.46 compare windows derived from the first vs. the last 40 VAS in texts. If text structure is meaningful, the trajectories should diverge. Of course in a shorter text, there is also more overlap between the two samples (first vs. last), such that with a median of roughly 70 across onDaF groups in CH, there is overlap between samples in more than half of the groups, while with a median of over eighty for three of the five BEL groups, there is no overlap. More diverging trajectories in BEL are therefore expected.

Figs. 6.45 and 6.46 show that there are systematic differences between the first and the

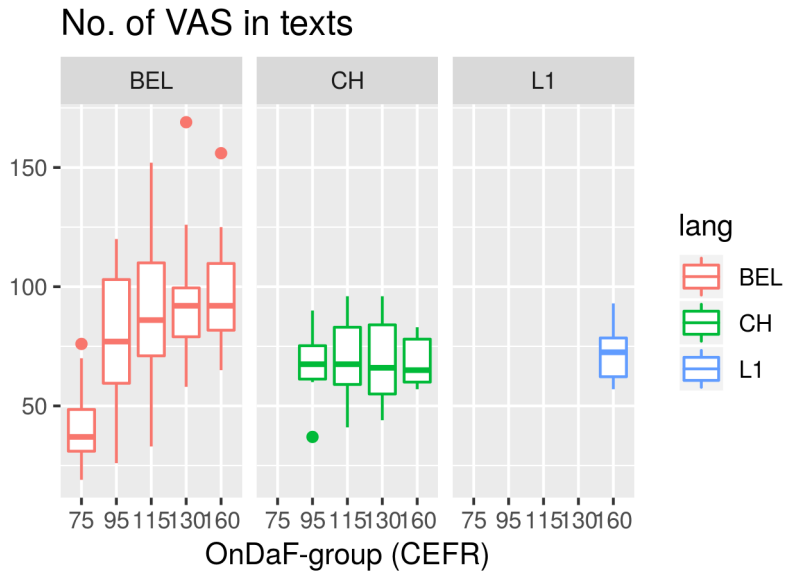


Figure 6.44.: Number of VAS in individual texts

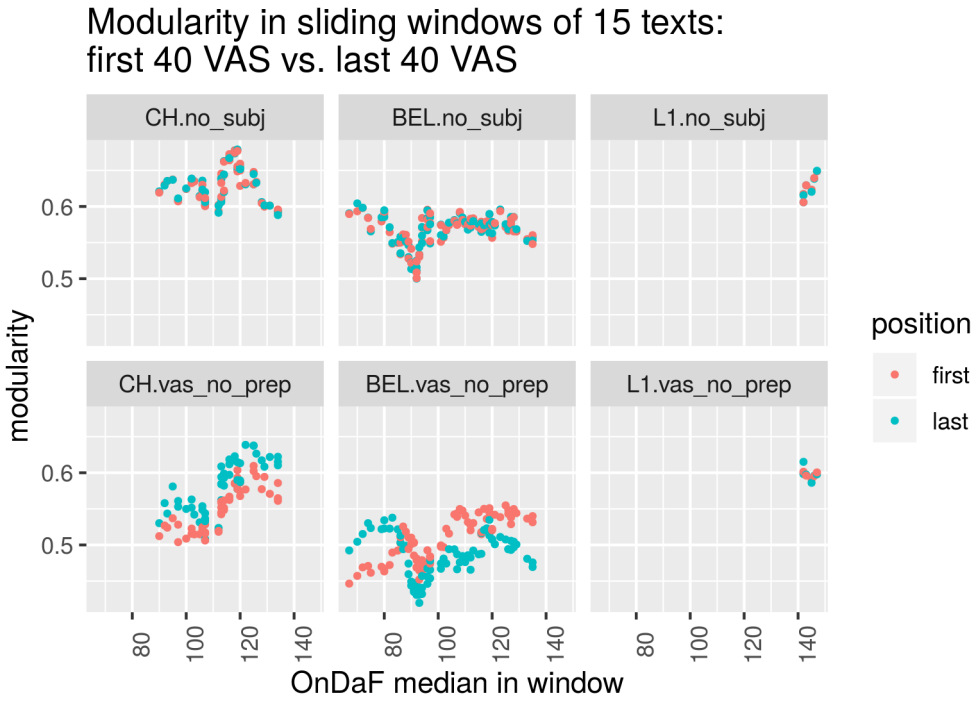


Figure 6.45.: Modularity in first vs. last 40 VAS in sliding windows (15 texts)

last VAS of texts in vas_no_prep, but not in no_subj:[23]

- In BEL, early intermediate show low modularity in the first, and high modularity

[23] Given the limited number of subject lexemes in L2 the latter is not surprising. The regression in the CH-no_subj graph does not seem to represent the data very well here, but the distribution is similar to previous analyses.

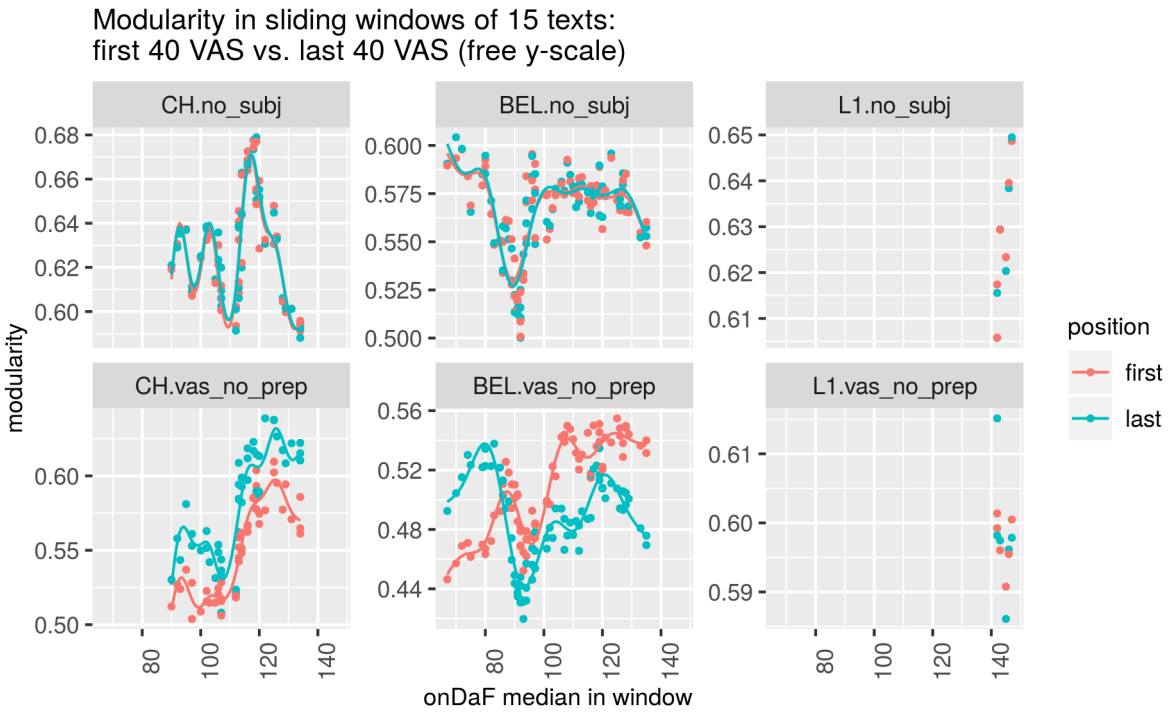


Figure 6.46.: Modularity in first vs. last 40 VAS in sliding windows (15 texts), free y-scale, with approximate trajectory

in the last 40 VAS of their texts. High-intermediate and advanced learners show the reverse pattern. In addition, very advanced learners' modularity does not drop in the first 40 VAS, but increases further, while it falls sharply in the last 40 VAS. Thus, the final drop in modularity seems to stem from text-structural effects.

- In CH, the trajectory of the last 40 VAS is almost exactly a copy of the first 40 VAS, only at slightly higher modularity. Thus, the final drop in modularity does *not* seem to stem from text-structural effects.

While trajectories are similar to previous analyses, this analysis reveals that modularity interacts not only with onDaF and text length, but also with text structure. Two explanations for this come to mind:

- One is rooted in progressive competence, where learners are aware that beginnings and endings *should* differ, or follow writing strategies such as an introduction followed by a differentiation or explication vs. examplification and connected conclusion. Such meta-awareness is part of language classes in middle and high schools, where students explicitly learn how to structure a text, that it should have an introduction, a central or main part, and a conclusion (*Einleitung, Hauptteil, Schluss* as it is taught in German classes). However, the L1 writers do not actually show this pattern, and in learners, it goes one way for CH, where the first 40 VAS are lower in modularity than the last 40 VAS, and the opposite way for advanced BEL writers (early intermediate BEL-learners show that first pattern). But this does not disprove the hypothesis, because L1-interference, cultural knowledge and preferred choice of register may all lead to inverse patterns grounded in existing meta-awareness. If CH, BEL, and L1 all

wrote in different genres or registers, even target-like meta-awareness could lead to different patterns if those belong to the different registers. A lot more research into text-structural and quantifiable ways of representing genre and register is necessary to fully understand this.

- A second interpretation lies in considering cognitive factors. Assuming that writing a text in a second language is something that requires some warming up to the language, where higher lexical differentiation and less randomness is reached at peak attention, this could explain the higher modularity at later text stages in early intermediate learners in BEL and very advanced learners in CH (because they need to activate the language and that takes a while), but with growing text length, attention goes down with cognitive fatigue, and texts get more repetitive and perhaps structurally simpler (more conceptually oral). Presently, there are not many studies into psycholinguistic effects visible in corpora, but this could be worth investigating further.

This section set out to answer two questions: Whether the final drop in modularity at high onDaF ranges in L2 observed in some of the previous analyses can be explained from text length effects, and whether BEL is comparable to the other language groups despite varying and growing text lengths. While a token-based normalization did not provide much clarity, based on the sample of 40 VAS from the beginning vs. the end of the text, text-structural effects become visible, and the final drop in modularity is weakened in vas_no_prep in BEL learners, but not in CH learners. The no_subj graphs show no differences between early and late text stages. This is odd given that the no_subj graphs are more variable in virtually all other respects, and that subjects are less variable in learners and thus taking them out should add variance. Clarifying this remains for future research. A text length and text structural comparison reveals that the absolute differences in modularity between BEL on the one hand and CH and L1 on the other are not caused by varying text lengths, suggesting that indeed, the groups can be compared.

6.4. Summary

The analyses performed in this chapter have shown that lexicosyntactic graphs based on verb-argument dependency exhibit clearly defined differences regarding the degree of their internal structure as represented by weighted Louvain modularity in interaction with the factors L1 vs. L2 and onDaF test scores in L2. Modularity is overall higher for L1 vs. L2 in vas_no_prep graphs, but not no_subj graphs, it is higher for CH vs. BEL, and shows a consistent drop at intermediate onDaF ranges in Belarusian learners for corpora $\geq$ 10 texts. Modularity in Chinese learners does not robustly decrease at intermediate stages in the same way, although a possible drop around the onset of the data in terms of onDaF scores is hinted at in some analyses. Chinese and Belarusian learners show marked differences in trajectories and distribution across all analyses.

Graph specificities between the full graph containing all lexemes and four categories of verb-specific graphs were shown to behave as predicted in aligning to developmental trajectories and higher modularity for specified graphs. Unlike predicted, the no_subj and the vas_no_prep graph types did not differ in the same way that vas_no_prep, vas_prep, and pp did. Results were shown to be robust, specific, and in line with most of the main hypotheses for corpus sizes $\geq$ 10 texts for the vas_no_prep graph, but much more variable for the no_subj graph, more strongly so in CH than BEL, and in L1. These differences

can partially be attributed to the smaller corpus size of the no_subj graphs, but likely point to a deeper grammatical and stylistic divergence between learner groups and even L1 authors.

In a first application of Louvain modularity to a research question in corpus linguistics, the measure has been shown to interact with corpus size and graph specificity in systematic ways. Convergence of modularity seems to be reached for the graph types presented here in samples of 20 to 30 texts in L1, and likely in less than twice that number in L2. This provides evidence to its applicability in small and medium-sized corpora.

To validate the method concerning the role of individual variance vs. corpus size and with respect to grouping and corpus sizes, two sampling techniques that are not commonly used in corpus linguistics at present were introduced, an out-of-sampling and a sliding window sampling. Both yielded results congruent with the initial onDaF-group-based analysis and the hypotheses presented earlier in this thesis, providing evidence that indeed, a) a grouping, and b) a grouping based a relatively simple assessment instrument such as the onDaF is able to produce reliable, consistent, and interpretable results representative of relevant aspects of a more continuous analysis even in a relatively small dataset. This is specifically relevant for any more qualitative research into the lexicosyntactic specifics of this data that cannot consider dozens of corpora for comparison but will need to rely on few, but representative groups.

It has been shown that a sliding-window-sampling is better suited for a more fine-grained analysis of intermediate stages than a grouping based on smaller onDaF ranges at least in an imbalanced corpus in terms of the distribution of texts across test scores. Corpus sizes of five or six texts have been shown to be least productive in providing robust results, even compared to the analysis of individual texts.

Results from a corpus size of 10 texts in the onDaF-based comparison were confirmed by all analyses of larger corpus sizes, and a corpus size of 15 texts in a sliding window analysis was sufficient for providing continuity of an implied trajectory. This can be seen as a first approximation of a lower bound for corpus sizes that yield usable graph metrics of this kind. Of course results first need to be replicated and usability of the measure confirmed on new data. A corpus size of one text in the individual text analysis shows onDaF and text length effects in line with the results from other analyses, but at such high modularity values that ceiling effects are likely to occur. A corpus size of 5 or 6 texts seems unfavorable, likely due to an interference of beginning emergent effects vs. leverage from large individual variance, while group effects stabilize and are overall larger than individual variance in 9/10 samples, so 9-text-corpora.

It can then be concluded that, in this specific corpus, subsamples of ten texts and larger despite not reaching the modularity limit relate reliably to samples based on language group and onDaF criteria and may be therefore be used for group comparison. In Kobalt, this means a comparison of either 10-text-samples over five data points in BEL and four in CH, as in the onDaF-based analysis at the beginning of this chapter, or a consideration of only three groups in BEL, but with 20 texts in each (BEL-95, BEL-115, BEL-130), and only two groups in CH (CH-115, CH-130), with 17 texts in each. The latter would still capture the u-shape in BEL, but for CH it would not represent the more interesting aspects of the trajectories as observed in the sliding window sampling. In fact, if one wanted to analyze the lexical specifics closer to the modularity limit in L2, it might be preferable to balance the CH dataset by collecting more data in the CH-95 and ideally also the CH-75 range.

Interfering effects of text length have been shown to be systematically related, more

so than their obvious influence on corpus size, to onDaF and text structure. It has been shown that a sampling based on verb argument structures from the beginning and the end of long texts is superior to a token limitation or text length normalization, likely because it is resilient to noise from other lexical and syntactic aspects and because text structural effects systematically occur at the beginning vs. the end. Text length and text structure effects occur in two ways, a) that texts of the same length authored by more advanced learners have higher modularity compared to those authored by intermediate learners, and b) that very advanced Belarusian learners start at high modularity and end at low modularity, while early-intermediate learners start at low modularity and end at high modularity; and that Chinese learners across onDaF ranges write at higher modularity at the end vs. the beginning of texts. This points towards either conscious changes in writing (for example, repeating thoughts at the end of the text for summary in advanced learners), or to unconscious, cognitively guided effects such as warming up vs. fatigue effects, or unplanned changes in register through the course of the writing process (effects from self-priming, for example). Text length effects have also been confirmed for L1 and CH, but not of the same impact, mostly due to lack of variance in text length in those groups compared to BEL.

The results reported are in line with the standard SLA assumptions in showing differences between learner groups by acquisition stage, between learners and native speakers, and a varying impact of individual variance, and provide evidence to the relevance of text length and text structural effects. A cross-validation across a number of factors shows that using Louvain modularity to compare lexicosyntactic structure, while not insensitive to these common influences, is robust enough to provide valid results even for an unbalanced corpus of limited size.

Remaining variance between samples, between learner groups and between L1 and L2 will be discussed in the next chapter in light of typological, cultural, and register-specific explanations. The discussion will also summarize and categorize some questions regarding future improvements of the model, both in terms of a better linguistic differentiation of syntactic, morphosyntactic, and lexicosyntactic categories, and in terms of the modeling of dynamic processes such as the emergence of lexicosyntactic constraints in individual speakers and in corpora, and the dialectics with other subsystems like syntax or text-linguistics.

7. Discussion and future research

This chapter concludes the discussion by first introducing some typological, cultural, and text-linguistic factors that may explain the remaining and unpredicted variance in the study. This is followed by a few suggestions to the extension and validation of the method, including a discussion of data size as a variable. Suggestions will also be made to the linguistic extension of the graph-based model to other linguistic research questions, not necessarily bound to the metric of Louvain modularity. Finally, results from the study performed in chapters 4–6 are tied back to the theoretical background presented in chapter 2, complete with a short sketch for an integration of coselectional preferences into usage-based linguistics in a functional rather than an epiphenomenological model.

7.1. Unexplained variance

The hypotheses in chapters 3 and 5, as they were derived from the presumed inner workings of interlanguage and previous theoretical and empirical work, predicted

a) higher modularity in L1 than L2;

b) higher modularity in advanced L2 vs. beginning and intermediate L2;

c) a u-shaped development of modularity in L2.

While a) and b) were confirmed robustly across analyses, c) was only robust for Belarusian learners.[1] What had not been predicted are two major differences between learner cohorts:

1. different levels of modularity, to the point where modularity curves of the CH-learners fit almost neatly between BEL and L1 for some subgraphs; and

2. different trajectories or progressions of modularity over onDaF scores, with the absence of a u-shaped trajectory in CH-learners.

This section offers possible explanations for the remaining variance from typological, cultural, and teaching perspectives, and a variationist point of view rooted in different choices of register and writing strategy. Of course these are only post-hoc interpretations. More research is needed to to gain more clarity concerning the dynamics and strength of those influences.

7.1.1. Typology

Typologically, Russian/Belarusian and Mandarin Chinese are on two ends of a spectrum, which is in part the reason they were selected for the Kobalt project: Russian/Belarusian are highly inflecting, synthetic languages. TAM is realized lexically to some degree, but

[1] Results from chapter 4 also corroborate the general tendency of an overuse of certain frequent coselections ("phrasal teddy bears") and an underuse of rare and more specific ones in learners until advanced learning acquisition stages, as it is described in the literature, cf. chapter 2.

largely through a rich verb morphology that encompasses not only tenses but also an array of productive prefixes to mark verb aspect. Mandarin Chinese, on the other hand, is highly analytical and lacks any verb morphology. TAM is marked lexically and lexicosyntactically through the modal (perfective or resultative) particle 了 (*le*). Syntactically, Mandarin is an SVO language with very limited flexibility in word order, while Russian/Belarusian have flexible word order tied in with information structure. Russian and Belarusian are partially and optionally pro-drop, while Mandarin is not. All of these aspects may play a role in coselection in different ways. However, two more features seem particularly likely to exert influence on coselectional constraint: Firstly, the presence of so-called verb-noun-compounds in Chinese along with the phonological feature of tonality and a high number of homonyms might cause higher contextual sensitivity in Chinese. Secondly, higher morphosyntactic complexity of the verb domain in Russian/Belarusian, i.e. a higher baseline of relationality expressed through morphosyntax in the verb domain, may in the absence of proficiency indirectly translate to an overall lower relationality of verb usage and thus lower coselectional constraint in the target language. A third aspect is not directly typological, but relates to the language environment of the BEL learners, that is their societal bilingualism. A higher number of competing concepts and lexemes and the dynamics of the bilingual mind and lexicon may lead to lower contextual sensitivity, resulting in lower coselectional constraint in L1 and L2. These three aspects will be discussed in the following subsections.

7.1.1.1. Contextual sensitivity in Chinese

Verb meanings in Chinese are often formed through verb-argument complexes such as 'look at' + 'book' = 'read a book' = 'read' (*kànshu*, 看书), but 'brightly' + 'study, read' = 'read out loud' (*lǎngdú*, 朗读); and 'run + step' = 'to run' (*pǎobù*, 跑步). These structures are sometimes referred to as verb-object-constructions, verb-noun- or verb-object-compounds, or inherent complement verbs (Bodomo et al., 2017; Badan, 2013). They are, as Bodomo et al. (2017) note, situated at the interface of lexicon and syntax. Some research suggests that some light verbs in Chinese are further undergoing a process of delexicalization (Cai et al., 2015; Xue, 2015), hence it is reasonable to assume that a learning strategy of Chinese learners in SLA lies in learning verb object combinations as holistic structures: If a verb is semantically bleached to the point of delexicalization in the L1, and the transparent part of the lexeme for an activity is the nominal part, a learner might prefer to rely on mappings of verb-noun complexes to verb-noun combinations in the target languages rather than learning verb senses and nouns separately as their dominant strategy.

At first glance, this may seem at odds with the observation that in Kobalt, CH learners use fewer, not more identical coselections than BEL learners in Kobalt (see section 4.1.2) – it would appear that if they had all learned the same mappings for the same concepts, this may not be expected. However, if one considers that BEL learners would be forced to productively recombine, choosing only from the most accessible parts of their vocabulary, while CH learners may have access to more differentiated, specialized, and ad-hoc retrievable mappings earlier on, a lower number of identical coselections may plausibly correlate with a higher degree of item-based knowledge.

Another aspect that may not only facilitate, but even require heightened contextual sensitivity in Mandarin is the large number of homophones. Mandarin has about 1 300 syllables including tone differences, but some 13 000 commonly used characters (Wiener et al., 2012; Chang, 1993). where each character denotes a syllable. Thus, on average, phonetic syllables have 10 separate meanings, and this refers to meanings as diverse as

认 *rèn* 'to know, to recognize', 任 *rèn* 'to assign, to allow' (and also 'office'), 妊 *rèn* 'pregnant, pregnancy', and 牣 *rèn* 'to fill up'. Many words are disyllabic, which can be used for disambiguation: Anecdotal evidence has it that speakers of Chinese habitually disambiguate by either "writing" the full character or the radical (the semantic core of the character) in the air or by asking "do you mean *rèn* from 责任 *zérèn* ('responsibility') or from 强韧 *qiángrèn* ('resilient')?".[2] I am not aware of how frequent or conventional this kind of disambiguation is in a native speaker context. However, a general sensitivity to coselectional constraints appears like a helpful strategy in an environment whose structure is as highly ambiguous.

It should be noted that this is not a statement about the status of lexicalization of these coselections in Mandarin. It seems entirely plausible, and likely, that they are fully lexicalized and perceived as words. However, the recurrence of those words in similar constructions still paves the way for generalization and analogical extension, i.e. speakers may notices that the same syllables (morphemes, lexemes) occur with other syllables (morphemes, lexemes) that may share some semantic features. This is plausible even where verbs are fully fused with the verb-noun-compound, as long as the nominal part is still accessible in the language – in the same way that German verb prefixes (such as *ver-* in *verbringen, verlassen, verkaufen* ('to spend (time)','to leave behind', 'to sell')) become available for analogical extension without ever occurring separately. Both contextualization and lexicalization could be facilitated through the morphosyntactic transparency of the verb-noun-compound.

7.1.1.2. Verb morphology in Belarusian/Russian and predication as a strategy

Belarusian and Russian form complex verb meanings through a range of prefixes (all examples are from Russian): *читать* (*čitat'*) – 'to read' vs. *прочитать (в слух)* (*pročitat'*) – 'to read through', 'to read out loud' (*v slukh*, literally 'into sound'), *перечитать* (*perečitat'*) – 'to read again', *подчитать* (*podčitat'*) – 'to catch up on reading', *дочитать* (*dočitat'*) – 'to finish reading', and many other variations.[3] It has been shown in chapter 4.2 that BEL learners use more prefix verbs than CH learners, and more even than native speakers in Kobalt. Particle verbs on the other hand are underused by both learner groups, perhaps even slightly more by BEL learners.[4] But how could complex verbs influence the degree of modularity between BEL and CH learners?

Complex verbs are semantically more specified than simplex verbs, which following Plank (1984) entails higher selectivity of their arguments. Learners from complex-verb-heavy languages should then logically show higher, not lower modularity. However, assuming Plank is right and complex verbs are more selective in Belarusian and Russian, too, the baseline for coselectional constraint might differ between the two learner cohorts in Kobalt: It is possible that in Belarusian and Russian, *only* complex verbs are prototypically coselectionally constrained, while simplex verbs are used with fewer semantic restrictions. In that case, the baseline for selectivity would be overall higher in CH learners in comparison, while coselectional constraints would only set in for more specified verbs in BEL learners, leaving them *less* coselectionally constrained as long as they do not use many complex verbs.

[2] Example only serves to illustrate the point, it may not be ideally chosen in terms of likelihood.

[3] A similar process exists with particle verb formation in German, which is a productive process, and even prefix verb derivation, which is also productive, albeit less so. Both are also known to be productive in learners of German SLA (Lüdeling et al., 2017).

[4] An underuse of particle verbs, but not prefix verbs, is also consistent with Lüdeling et al. (2017).

This point relates to the quality of verbs and their arguments in learners at various stages in Kobalt. It has been discussed in chapter 4.2 that V+OBJA and V+OBJP combinations in particular are functionally and structurally different at different acquisition stages and in learners vs. native speakers. Learners at lower intermediate stages overuse semantically light verbs like *haben* ('to have') or *geben* ('to give, to exist'), while at the most advanced stage, verb and argument coselections of a different kind begin to appear, viz. support verb constructions or semantically richer coselections (such as *Arbeitsplatz finden* ('to find a job') or *unter Hunger leiden* ('to suffer from hunger, to starve')).

Light verbs are not just simpler verbs without further linguistic repercussions. They may in fact reflect a differently weighted grammar altogether, one that is centered around predication rather than relational expression, as would be the case for more complex verb-argument structures (of which there are fittingly few in Kobalt-L2, see section 4.1.5). Predication is not limited to adjectival predicates attached with a copula, but can be semantically or syntactically more or less complex.[5] The difference to complex verb argument structures is not necessarily in the syntactic, but in conceptual complexity or simplicity, the number of active concepts or agents, and their relational density. Light verbs facilitate a predicational syntax through adding semantic aspects to the predication while also lowering cognitive load in the command of their morphosyntactic paradigm, since they are frequent verbs in various functions as auxiliary, modal, or lexical verbs with different semantics as in the case of German existential *geben* ('to give, to exist'):

> "The function of light verbs is to modulate the event predication of a main predicator in the clause. Different light verbs will do so in different ways and some of the semantic contributions are quite subtle. This is in part because of the flexible interpretation of the underlying lexical semantics. The verbs which allow light verb readings have lexical semantic specifications that are of a very general nature. This allows them to appear in a variety of syntactic contexts. The idea that light verbs and their corresponding main verbs are derived from one and the same underlying representation accounts for the fact that light verbs are always form-identical to a main verb counterpart in the language (...)" (Butt, 2010, 74).

Haben 'to have' and *geben* 'to give, to exist' might be used as frequently because they are semantic prototypes for arranging a predication rather than denoting processes – they are in fact used very similar to the copula *sein* 'to be', only with an additional existential or possessive extension. Russian/Belarusian allow for the omission of 'to have' in a possessive context in the same way that copula verbs are not required. While it is possible to express the proposition 'she has a diploma' through the use of the equivalent to 'own' (*иметь*, *imet'*: *Она имеет диплом*, *ona imeet diplom*), it is rather unidiomatic compared to the verbless *у неё диплом* (*u neyo diplom*, literally: 'at/with her diploma') or the existential *у неё есть диплом* (*u neyo est' diplom*, literally: 'at/with her is diploma'). German existential *geben* ('to exist') can be described similarly. Thus it might be worth considering that early intermediate BEL learners do not use verbs to denote complex processess as frequently, and thus do not access or do not require access to coselectionally constrained areas of the semiotic space as frequently or strongly as more advanced learners do.

[5] Müller (2002) provides an in-depth discussion of complex predicates for verbal complexes including modal infinitives, resultative constructions, and particle verbs, all of which are underused in Kobalt-L2.

This may be related to a structural shift in the development of syntactic complexity in SLA,[6] where the use of language may not differ gradually, but categorially and imply wider-scale repercussions on other linguistic and cognitive aspects. In a sense similar to how predicate logic extends, and also differs from, propositional logic, this would also map back to an idea outlined in chapter 2: Early (intermediate) learner language may be better described through more general cognitive processes (like mapping predicates), in line with Klein (1998)'s notion of learner varieties, while more advanced L2 might better be described relative to an interlanguage space that is shaped by cognitive aspects, within-system emergence, and most importantly the three languages involved: The L1(s) of the learner, the target language as perceived by the learner (=latent structure in Selinker (1972)'s terms) and the target language as it exists in the learner's input.

While this is certainly an interesting aspect to look into in the future, let it be said with Occam's razor that holistically learned V+NP combinations in Chinese learners would provide a simpler explanation for differences in modularity values, and also the lack of a u-shape. It would, in fact, mark the difference between a u-shaped development in a distributional learning scenario, and a more or less linear growth in an item-based scenario, as these were discussed at the beginning of chapter 3. It is possible that a combination of typological and teaching or learning strategies materializes in a more item-based development of coselectional constraint in Chinese learners, while the u-shaped development in Belarusian learners stems from a sudden randomization and then a reassembling into new structures.

That would, however, mark a surprisingly large difference between the two language groups if interlanguage can be considered somewhat comparable at an equivalent onDaF score, or in other words, if L1-independent language assessment of L2 is linguistically valid. It seems that at least *some* process of randomization should still be visible even in a more item-based acquisition process, if hypotheses were right about a clash between an existing, relatively fixed interlanguage system, a necessity to succeed in much more complex communicative situations, and combinatorial aspects in an emergent system. If this whole problem could be avoided by a simple strategy of learning V+NP combinations, it would not only have greater pedagogical implications: The necessity of temporary loss of accuracy, or u-shaped development, for the process of acquiring a system that contains both general rules and a number of idiosyncrasies, as suggested in Carlucci and Case (2013), would be called into question, and with it the structural interpretation of coselection acquisition and the role of general, as opposed to typologically determined, cognitive mechanisms on language learning. If Chinese native speakers show higher sensitivity for coselectional constraints earlier than native speakers of other languages, then this should have wider repercussions and be detectable in other aspects of their FLA and their SLA as well. Rather than in the structural approach in this thesis, an item-based development can only be tracked in an item-based fashion to ascertain that it is indeed item-based in a meaningful way and not random. This most likely would require extensive true-longitudinal data.

In an interesting parallel, it has been shown that children have an easier time acquiring verbs vs. nouns at a certain stage in FLA (Imai et al., 2005; Tardif et al., 1997; Tomasello et al., 1997; Goldfield, 2000) – in most, but not all languages analyzed so far, and specifically *not so* in Mandarin Chinese (Tardif, 1996; Tardif et al., 1997). The question of

[6] See Ortega (2003) for a synthesis of 27 studies from the 80s and 90s, Vyatkina (2015) for a more recent overview.

whether this noun bias in children reflects a developmental constraint or a linguistic feature of their L1 is also raised by Ortega (2013, 12) as a question for different settings in SLA research. Could it be that Mandarin and Belarus/Russian provide such fundamentally different frameworks for learning both verbs and verb-argument coselection that they entail majorly diverging processes in target language construction?

The task for future research would then lie in modeling expectable differences and comparing them across language pair matrices (German, English, Mandarin as L1 and L2). If it turned out that this difference is a common aspect of the interlanguage development at least with some language pairs, there would be two major questions for an improved theoretical model of SLA:

- Is a u-shaped development *structurally necessary*, as Carlucci and Case (2013) suggest and as has been discussed widely in FLA research? Or is it reflective of a preference or a certain strategy, which in SLA is sensitive to L1-transfer?
- Is advanced learner language *relationally and semiotically* different from early learner language, with a shift from intermediate to advanced learner German in Belarusian learners marking also a shift into a subspace of interlanguage where coselectional constraint gains relevance?

While all of those interpretations may seem – and are – rather speculative, the following two examples serve to illustrate the degree of divergence between the two cohorts that requires explanation. The first text is written by a Chinese learner, the second by a Belarusian learner. The texts were chosen at random (not selected for being particularly good examples), but by proximity in onDaF. They lie apart by only one onDaF point (87 vs. 86), at which both authors can be attributed to early intermediate proficiency levels.

(a) "Die meisten Arbeiter gingen nicht in die Schule, deshalb informierten sie sich wenig über die Welt und die Technik und Wissenschaft. So wurden die Güter von ihnen nicht gut hervorgebracht. Die meisten Arbeiter waren jung. Sie müssten den ganzen Tag arbeiten und die Arbeitsbedingungen waren schlimm. Nicht nur die Arbeiter, sondern auch die meisten Jugendlichen sowie die Erwachsenen führten ein schlechtes Leben."

'Most workers did not go to school, that is why they did not enquire much about the world and technology and science. Thus the goods were not produced by them well. Most workers were young. They would have to work all day and the working conditions were bad. Not only the workers, but also most adolescents as well as the adults lead a bad life', (*CMN_057*).

(b) "Und wer ist die Jugend? Es sind die jungen Mädchen und Jungen, die viel Energie, Kräfte, Willen und Träume haben. Die Jugend studiert sehr schnell, akzeptiert die neue Information leicht. So ist es immer: Wie die Leute sind, so ist die Welt oder wie die Zeiten sind, so sind die Menschen. Also waren die früheren Generationen ganz anders als die heutige."

'And who is the youth? It is the young girls and boys, who have a lot of energy, forces/strength, willpower, and dreams. The youth studies very fast, easily accepts the new information. It is always this way: How the people are, such is the world or how the times are, such are the people. Thus the previous generations were completely different from today's', (*BY_033*).

My personal intuition suggests that, in the Chinese text, the following coselections are not only acceptable, but highly natural in German:

> *meisten + Arbeiter* ('most' + 'workers'), *gingen + in die Schule* ('went' + 'to school'), *informierten + über + Welt* 'enquired' + 'about' + 'world', *Technik + Wissenschaft* ('technology' + 'science'), *Güter + hervorbringen* ('to produce' + 'goods'),[7] *den ganzen Tag + arbeiten* ('all day long' + 'work'), *Arbeitsbedingungen + schlimm* ('working conditions' + 'bad'), *nicht nur + sondern* ('not only' + 'but'), *die Jugendlichen + sowie die Erwachsenen* ('adolescents' + 'as well as adults'), *führten + Leben* ('lead' + 'life'), *schlechtes + Leben* ('bad' + 'life').

In the Belarusian text, I personally find the following coselections highly *un*idiomatic in German:

> *junge + Mädchen und Jungen* 'young' + 'girls and boys' ('junge Mädchen' sounds antiquated in this context, while 'junge Jungen' is phonotactically strange); *Energie + Kräfte* ('energy' + 'forces'), *Willen + Träume* ('will' + 'dreams'), *Jugend + studiert* ('youth' + 'studies'),[8] *akzeptiert + Information* ('accept' + 'information'), *akzeptiert + leicht* 'accepts' + 'easily'; *wie + Leute + sind + so* ('how' + 'people' + 'are' + 'thus'),[9] *wie + Zeiten + sind* ('how' + 'times' + 'are'), *heutige + Generation* ('today's' + 'generation').

Obviously, my intuition may not be reliable. For a stronger argument, frequencies of all of these coselections should be collected from DeReKo (Leibniz-Institut für Deutsche Sprache, 2019). I would still argue that the difference is both qualitatively and quantitatively so large that it cannot be explained from different *conscious* strategies or learned behavior from different teaching traditions; which makes specific guiding or constraining forces on the Chinese learners' side appear plausible. For a better answer, a comparison with other language groups, other target languages, and experimental research looking into context sensitivity in interaction with L1 is required.

From the analysis in this thesis it cannot be decided whether either of the learner cohorts marks the regular case or if perhaps both are a deviation from a norm, or whether such a norm even exists: It is possible that coselectional constraint is generally more strongly developed in Chinese-speaking learners of German, or that it is generally less strongly developed in Belarusian/Russian-speaking learners or German, or both. This can only be clarified in a triangulation with data from further L1 groups.

7.1.1.3. Belarus as a bilingual language environment

This final aspect is not typological, but concerns the language environment of the BEL learners. Belarus is a bilingual country, where all school students are immersed in a bilingual educational system from age 7 and language contact is widely present at school and in professional and everyday life. With Russian as the dominant language in educational and professional contexts, Belarusian tends to be more on the receiving end of language

[7] This one appears semantically unusual in the context, but not coselectionally strange. In fact, it even occurs once in the German reference corpus DeReKo (Leibniz-Institut für Deutsche Sprache, 2019).

[8] It should be 'Jugendliche studieren' or 'junge Leute studieren', 'adolescents' or 'young people + study', although in German culture, students are not typically categorized as adolescents, and rarely identify as adolescents.

[9] There is a conventionalized phrase, *wie die Leute halt so sind* ('that's just how people are'), but the use in this example seems to differ both pragmatically and semantically

mixing and change, but some reverse influence has also been observed (Zaprudski, 2007; Hentschel, 2014; Kittel et al., 2010).

It is known from the study of bilingulism in general that it affects lexical retrieval in both languages (Sandoval et al., 2010; Bialystok et al., 2008; Ivanova and Costa, 2008; Gollan et al., 2008; Khateb et al., 2017; Prior and MacWhinney, 2010). With two languages more or less constantly activated in environment of the BEL learners, it is likely that a lot of transfer between the two languages happens, and some unintentional code-switching may be part of this. Leshchenko et al. (2018) describe a similar process for speakers of Russian and Komi-Permyak, a Finno-Ugrian language spoken west of the Ural Mountains with little lexical overlap with Russian per se. The authors suggest that "a "fused" zone of syntactic and lexical representations in [the] bilingual mental lexicon provides the basis for extensive unintentional code-switches in bilingual speech" (Leshchenko et al., 2018, 301). Their analysis is based on a word association experiment with expert speakers of both languages (native speakers training to become school teachers in both languages) in which for a word given in one language, up to 69% of the freely associated words were of the other language. Participants also rated most of these bilingual adjective+noun, verb+adverb, or verb+object combinations, e.g. 'listen attentively', 'be on time', or 'native language', as frequently used and heard.

While the matter of code-switching is well-researched in linguistics in general, the effects of bilingualism – a higher combinatorial power, more lexical association beyond the conventions of each language – on coselectional constraint – the limitation of combinatorial power by specific convention of a single language – do not seem to have attracted much research so far. There is some work on learners in an English as a lingua franca (ELF) environment,[10] but these cases are different. They look at speakers who have not fully acquired the idiomatic conventions of the spoken language, which does not apply fully acculturated bilinguals in a constant immersive bilingual environment. It is known of course that language contact is a driving force of language change. It is likely easy to find some coselectional preferences that are higher in Belarusian Russian vs. Russian in Russia, but that is not the same case either: The question relevant to this study is whether bilinguals in a fully bilingual environment may overall be less coselectionally constrained. In other words, do they use and recombine words more flexibly than monolingual speakers, and if so, does that also project to their second languages? While I am not aware of any studies into this, it could be a particularly interesting field of study for a clarification of language-specific vs. general cognitive mechanisms in the research of coselectional constraint.

7.1.1.4. Summary: Typology and language environment

In summary, differences in typology and language environment offer one strong and one tentatively plausible explanation for the unpredicted variance between groups: Verb-noun-complexes may be more likely to be continuously learned as holistic items in CH learners, leading to a lower degree of randomization or breaking up of holistically learned structures from early SLA. This could explain both a higher degree of modularity and the lack of a u-shape. In BEL learners, the societal bilingualism of their environment might contribute to more competition between lexical and syntactic forms and competing constraints leading to higher perceived noise in the input, and expressed in lower contextual sensitivity in

[10] For example Kecskes (2015) on whether the idiom principle might be blocked in L2, or Pitzl (2012) on creative uses and remetaphorizations in ELF.

bilinguals in general. This could explain the drastically lower modularity values and a higher willingness to randomize in BEL learners.

Both of these hypotheses could be tested relatively easily through a triangulation with learner data from other L1 environments. The simplest with respect to controlling for linguistic environment would be to add data from monolingual Russian speakers, who are culturally and linguistically close to Belarus, and who could shed light on the role of the bilingual environment. A comparison with other L1s which neither have a rich complex verb morphology, nor a high rate of verb-object-compounds might give insight into the question of the role of item-based vs. distributional or randomization-based development. In any case, a triangulation with several other L1s would serve to clarify whether the CH or the BEL results are the more regular case; if such a norm exists; whether there are two – a high vs. a low – modularity groups or more, or perhaps a full spectrum or continuum of typologically powered in- or decreases in modularity; and whether learning trajectories can be clustered systematically at all. The question of whether a u-shape would be observed in CH in lower proficiency data, but is missed in Kobalt due to the late data onset, could be answered by completing the Kobalt data with another cohort of CH learners. Unfortunately, this might introduce new artifacts from topic effects, since nearly a decade has passed after the initial collection and new participants would likely write different, and differently themed texts now. An assessment of the comparability of modularity values in older vs. newer texts and of topic-mixed corpora in general would be required first.

7.1.2. Register, cultural, and teaching effects

It has been pointed out several times in this study that the three cohorts in Kobalt seem to follow different writing strategies and generally seem to produce texts from different registers in response to the same prompt.

Education systems both in China and in post-soviet states are known to place an emphasis on rote learning and the near-identical reproduction of previously heard or read input. This might suggest that, while concrete lexical realizations may be unpredictable, learners would have a template ready for how to respond to argumentative essay prompts in terms of text structure, discourse positioning, and other factors. While indeed such a tendency exists for each of the groups, between all three cohorts, essays differ remarkably in their larger textlinguistic aspects and in the lexicosyntactic details; and not all of those aspects appear likely to stem directly from teaching effects or be desirable in classroom text production.

Differences between L1 and L2 cohorts may in part be attributable to the younger age of the L1 participants, who were high school students at the time of data collection, while the L2 data was collected in 2nd–4th year classes at colleges in the L2 countries. Despite the aim of the Kobalt project to compile a deeply homogeneous and controlled corpus, this was somehow not considered problematic. High school and college students, in spite of their proximity in age, differ not only regarding their education levels, but also represent a different group selection. German high school students in *Grundkurs*, the lower of two self-chosen levels of German classes in high school, are not necessarily a group of self-selected language enthusiasts. Much unlike the L2 participants: Studying a second language at college major level, and not English as the most prioritized L2 in the world today, the learners in Kobalt would perhaps best be described as counterparts to what in Germany would be university students of French or Spanish. Those would likely produce more sophisticated texts in terms of structure and content compared those in Kobalt-L1.

Relevantly, the small difference in age – L1 participants were typically 17, learners around 20 years of age – marks a large difference in their self-perception in relation to the prompt: L1 writers largely view themselves as part of the generation to be evaluated, while learners discuss adolescents or youth as a remote group. It is surprising that this was not reflected upon in the data collection planning for Kobalt.

Yet looking through the data, it would be a stretch to read clear argumentative advantages into the learner texts per se, despite their higher level of formal education. Rather, it appears that writing choices are culturally shaped. It seems that Chinese and Belarusian learners of German, and German high school students of 2012, all reply to different, culturally shaped expectations of how to answer the question whether previous generations had a better life: Belarusian learners tend to agree, Chinese learners tend to disagree, and native speakers tend to question the validity of the question itself and suggest that the answer depends on which generations are compared with the current ones. While there is some diversity within groups as well, this is only a slight simplification of what appear to be culturally and historically shaped answers to a culturally charged prompt.

Would a better, less charged prompt have yielded more comparable results? The problem is that the prompt needs to be motivating and engaging in order to elicit results at essay length. At the same time, it needs to presume as little contextual knowledge as possible *for anyone to be able* to create such an essay-long answer without preparation or additional material. So how does one create long answers without prior knowledge or prepared material? Likely through implicit knowledge, which is cultural bias.

Cultural bias does not only shape the response content, but also its contextualization. Chinese learners, for example, frequently refer to a supposedly existing cultural discourse on the matter:

> (1) “Wenn ich vor dem Computer sitze und im Internet surfe, finde ich einige Artikel über den Vergleich zwischen den Jugendlichen und früheren Generationen. Manchmal höre ich die Anrede “80er” oder “90er”, die eher eine negative Bedeutung den Jugendlichen verleiht”, (*CMN_017*).
>
> ”When I sit in front of the computer surfing the internet, I find some articles about the comparison between adolescents and previous generations. Sometimes I hear them addressed as “80s”or “90s”, which gives adolescents a rather negative meaning”, (*CMN_017*, rough translation).
>
> (2) “In Bezug auf die Frage “Geht es der Jugend heute besser als früheren Generationen” führten die Leute eine Diskussion und verschiedene Antworten wurden gegeben. Manche Leute stellen auf die Seite, dass es der Jugend heute besser als früheren Generationen geht. Infolge sind ihre Argumentationen dazu da. Zunächst meinen sie, dass die neue Generation einen weiteren Horizon erhalte”, (*CMN_051*).
>
> ”Regarding the question “Are young people today doing better than previous generations” people had a discussion and different responses were given. Some people put on the one side that young people are doing better than previous generations. First, they think that the new generation has a wider perspective”, (*CMN_051*, rough translation).

While such reference to authority or 'objective' facts can reflect an individual hedging strategy, it occurs so frequently in CH writing, but not in the other cohorts, that it seems more plausible to assume that it is rooted in a cultural expectation. This is in line with Wan (in prep.)'s analysis that finds systematic differences in the argumentative structure between L1 and CH texts in Kobalt. This is an example of a bias that can be explicitly taught and learned – "refer to the authority of the group to solidify your argument".

Another cultural and teaching aspect that may be taught and learned explicitly is the question of what it means to discuss a prompt. Am I supposed to take a side? How strongly so? Am I supposed to write vividly, engaging to the reader, or matter-of-factly? The hyperreferentiality of discourse implies that asking for an answer to the prompt does not just elicit an answer to the prompt, but also a contextualization of the learner or native speaker in relation to the prompt, the answer, and the cultural discourse. In answering a question I inescapably define my own role and with it the relational environment. This includes the question of whether a participant thinks they put themselves outside of what they suggest is the domineering cultural discourse on the matter and is part of what creates register choice. Such contextualization may sharpen over the process of writing, as in the example that was provided in section 6.3.4, where a BEL learner starts out rather matter-of-factly and analytically, and shifts to a more emotionally charged register mid-text. This then is unlikely to have been learned and taught explicitly, but likely reflective of cognitive and emotional process in the individual writer. It is not unaffected by cultural bias, though:

The German expression *sich in Rage reden* ('to talk oneself into a rage') refers to speech that gets more emotional and divisive through its course. In Kobalt, this can be interpreted as happening frequently in the BEL texts, but not the CH or the L1 texts. It also correlates with the longer texts of Belarusian learners with growing onDaF, and could be interpreted as their 'giving it their all' – writing as much as they can, with full personal engagement. But of course this does not occur in a thematic vacuum, but correlates with a cultural narrative in post-soviet countries, where the collapse of the former system has left an overall destabilization and a certain nostalgia for the more structured, if less consumerable, soviet times. A black-and-white narrative, in which things are sharply, and often overly, contrasted is congruent with a spirit of nostalgia. It is much less congruent with an optimistic narrative. Thus the choice of one side in a debate where the two sides are of such different emotional value may be equaled with the choice of a thematic, register, and lexical path. This has repercussions on the coselectional properties of the text, too.

In addition, self-positioning in the space of a more divided discourse is likely to prime rhetorical and linguistic concepts that would not be primed in a less charged discourse space. For example, in German society, there used to be the trope or a cliché discourse *Opa erzählt aus dem Krieg* ('grandpa talks about the war'). While nowadays, most grandfathers in Germany are from the post-war generation, for at least one generation, this referred to a discourse frame where certain slots were opened for the listener and/or interlocuter, such as the annoyed family member who has heard enough of the old stories, or the appeasing family member who wants grandpa to be able to tell his story despite having heard enough of it, the empath, the fascinated child, the judge of war crimes, and so on.[11]

Presuming that a similar discourse might exist in the Belarusian society today about soviet times vs. post-soviet times, participants familiar with this discourse will likely

[11] This example only serves to illustrate the principle, it is not intended as anything beyond a very superficial description of the role of the communcative system in discourse.

be primed for its whole frame and with it linguistic aspects of it, but in interaction or interference with second language thinking and expression. Chinese learners and native speakers seem much more detached from the topic, but they also overall agree more on things being better now – thus their detachment may prime certain linguistic structures, but the cultural discourse may prime their detachment in the first place.

Native speakers and CH learners are also more homogeneous in their essay structure, which is likely a teaching effect: They either present a thesis or ask a question first, then bring arguments and counter-arguments, and then conclude with a final statement of how they see things. This corresponds to the genre of *Erörterung*, a strictly structured debate-style argumentative essay that German students are taught in middle school, and in which analytical reasoning and personal detachment is expected.[12] CH texts are further rather critical of some aspects of Chinese policy (many critically mention the Cultural Revolution or the one-child-policy), suggesting that learners are sensitive to the expectation of critically assessing and evaluating a situation, including the larger, more abstract powers affecting it, in an argumentative essay of this kind. It is equally plausible and easy to remain vague or completely step away from a political assessment, as the example of the Belarusian learners shows: While they frequently discuss different states of the society at large, those are never tied back to a political debate in the country, but rather exemplified with stories from the background or family history of participants. The degree of willingness to share personal stories like these, and what to include in those, depends not only on the teaching of text production in schools, but also on the more general cultural expectations, value systems, and categorizations around personal and public space, intimacy, and rationality. In this wider context, coselection appears like a function of many parameters beyond the open or idiomatic choice of words.

There are additional factors influencing the writing process outside of the scope of a purely linguistic analyis. Some experimental evidence suggests that reasoning is affected by the use of either L1 or L2 (not a specific language) in judging the best way out of the trolley problem: A loose trolley (train wagon) is running towards a group of five people who are tied up on the tracks. Using a lever, the trolley can be redirected to a sidetrack on which only one person is tied up. The question is, will a participant agree to pull the lever to sacrifice one person in lieu of five? Or will they refrain from actively manipulating the situation? The first is a utalitarian solution, whereby the active involvement in killing a person is considered moral if it is for the cause of saving the lives of several. Costa et al. (2014) find that in groups of English L2/Spanish, Hebrew and French L2, and Korean L1/English L2 speakers, a strikingly higher number of speakers choose the utalitarian path in the L2-groups compared to the L1-groups (7.5% vs. 0 in the Korean-English group, 44% vs. 28% in the English-Spanish group). If judgment and reasoning are affected so strongly by use of L1 vs. L2, then it is reasonable to assume that answering the same prompt in a first vs. a second language does not follow the same mechanisms, and it is consequentially also reasonable to expect linguistic differences that are not only due to lacking L2 skill, but also a reflection of a different kind of processing.[13] Veltkamp et al. (2013) even found

[12] Prompts are typically student-centered, prototypical examples include "should students wear school uniforms?" or "should students be allowed to bring their phones into the classroom?".

[13] Costa et al. (2014) name a lack of emotional or sentimental processing as a reason for the differences between the L1 and L2 groups. This refers to a discourse in experimental philosophy where moral judgments are discussed as an interaction of rational and emotional choices. However, in relation to Kobalt, this does not seem plausible, since many of the BEL texts on the contrary appear rather emotional.

systematic differences in participants' results on the Big-Five personality test (openness, conscientiousness, extraversion, aggreableness, neuroticism), depending on which language it was taken in (German vs. Spanish).

It seems, however, that in the most advanced learners, who score higher than most of the native speakers on the onDaF test, texts assimilate to native speaker texts not only in linguistic details, but also in more general tone and register, and become less emotional in Belarusian learners. Congruently, Čavar and Tytus (2018) find no L1/L2 effect in an extension of the trolley problem experiment to a group of fully immersed bilingual speakers of German and Croatian. They map this to a higher degree of enculturation in both languages, so that none of the languages is perceived as a second language, and that both are rooted in social interaction, more so than in the L2 groups in Costa et al. (2014). While even the most advanced learners in Kobalt would likely not reach this level of enculturation in German, the effects on reasoning and emotional processing may very well be tied in with cognitive load and spontaneous mapping of evalutative categories such as 'good' or 'bad', and higher differentiation may require more language-specific and available cognitive resources. In that sense, it is possible that a more balanced, more rational, and less emotionally charged text production is not simply a matter of choice, but a result of a complex function that includes aspects from priming, cultural bias, cognitive resources, interaction of concepts with nervous system functions, and so on.

How is all of this relevant to the discussion of coselectional constraint in Kobalt? Firstly, while all of these observations are interpretative at this point, some of the remaining variance in Kobalt may be explained through culturally shaped responses and teaching effects (presumed reader expectations, writer positioning, delivery of arguments in more or less entertaining or engaging ways), as well as L2 effects on reasoning, especially where the L2 is not yet easily processed.

Secondly, it was suggested at several points that differences between texts do not necessarily stem from cohort specifics per se, but from the cohort-specific register choice: Perhaps BEL learners are not *per se* less coselectionally constrained, but only choose registers that *make them appear that way*. This of course raises the question of comparability: How can texts of different registers, written by writers of (subjectively) different ages and at least two levels of education be meaningfully compared? Can they even be meaningfully compared at all regarding their linguistic details?

With the phenomena of culturally shaped discourse embedding and L2 effects on reasoning, it seems that there is no way around this anyway: If no contextualization is provided to guide participants into a specific answer to a question, they are bound to react from their culturally rooted standpoint. This is the plain reality of second language. Language learners are not simply less proficient speakers of a target language, they use a second language from the embedding into another cultural and linguistic system, unless or until they are acculturated to the target language environment – similarly to how children are not simply little native speakers with lack of control over certain syntactic areas, but are cognitively and socially unlike adults in many ways.[14] While some organizational principles of interlanguage may be general and systemic, target language *in use* will likely always be shaped more by cross-linguistic and cross-cultural aspects.

Are texts in Kobalt still comparable despite their differences? Yes and no. First of all, anything can be compared, the question is by which feature and what kind of learning a comparison by that feature might generate. Of course it is fair to say that coselec-

[14]This is not meant to equate language learners or less proficient speakers with children in any way.

tional preferences will take different concrete realizations in a more narrative vs. a more evaluative text. Comparing those concrete realizations then is a dead end, as has been shown empirically in chapter 4. But if in both the more narrative (BEL) and the more evaluative (CH) texts, coselectional constraints exist in different ways in interaction with onDaF criteria, this is indeed a fair comparison. It is also simply the case in this observation that, given the same outward conditions and means, learner groups and L1 writers differ systematically in text linguistic dimensions higher than syntax. This is an aspect of learner language in use and thus relevant to consider in usage-based SLA research.

However, if with this knowledge one set out to collect new data, would it be worth considering a "better" prompt? "Better" is written in quotes, because creating one is really not a simple task. As argued above, the prompt needs to be engaging and interesting, but answerable without additional material, which will then necessarily draw from the participants' cultural biases. It also needs to be one that does not lend itself to a lot of text reuse or priming. Studies have shown that categories as abstract as German particle and prefix verbs can be primed through a prompt: If the prompt contains *any* particle or a prefix verb, not only the lexeme, but the whole category will occur more frequently in the response texts (Lüdeling et al., 2017, footnote 18). In Kobalt, this is reflected in the frequent reuse of *gehen* ('to go') in its constructional sense of 'to be (well/unwell)', and the frequency of prompt-related words like *Jugend, Jugendliche, Generation* ('youth', 'adolescents', 'generation') is certainly grossly exaggerated compared to the frequency of occurrence in the natural production of the same speakers. In fact, one might argue that adolescents rarely, if ever, refer to themselves or their peers as *Jugendliche* or to people of different ages in their family as *Generationen*. Since any prompt will contain syntax and semantics and prime for those, some of this is unavoidable.

The question then is not whether elicited texts may be compared to one another – they may, even if some of these comparisons yield negative results because no two similar items can be found. The question is rather: What is the ontological status of the elicited data? If a data collection sets out to collect comparable data, what is it trying to compare? The texts here can be understood as a mapping of the function of the prompt with the learners' response system as an input variable:

$$f : \textsc{prompt}(\textit{learner response system}) \mapsto \textsc{text}.$$

If certain aspects of the prompt map to culturally specific aspects in the response system, those mappings will be visible in the response. If two groups of learners have structurally different response systems (interlanguages), the function of the prompt will map to different texts. The prompt is, as it seems, not a sampling function, one that will give balanced relations of what the response system is capable of with other functions. A learner text is not a neutral sample from their language production, it is a sample of their language production as funneled through a specific filter, and measures should be interpreted in this context.

One of the main decisions in corpus compilation is then: Do I want to build a zoo? Or do I want to build a parliament? In a zoo, exemplars are of interest. For Kobalt, the concrete realizations of language in a single instatiation (data collection) by different speakers are like an animal in that zoo. No-one would recommend representing animals relative to their natural number of occurrence in a zoo: billions of mice and a third of a polar bear. It is interesting to look at different types of bears, perhaps compare their sizes and shapes and colors, but not how many of each there are.

In a parliament, *only* relative representation matters. Translated to learner corpus research, a parliament would require representation of the most frequent learner language,

which is likely not reached by argumentative essays on most levels of acquisition. In other words, Kobalt is decidedly *not* a parliament, it is not representative of Belarusian or Chinese learner German *in general*, and it is not representative of coselectional constraint in those two groups or L1 German either. It is the output of a mapping function of a specific prompt in which for a large number of lexemes and lexicosyntactic structures, overlap exists between all three groups – thus suggesting that the input systems are related. But it also shows that despite these similarities, the ensuing textualization is still very different. The research question defines whether representation relative to the total distribution is necessary, or whether exemplars provide sufficient richness and density of information. A consideration of these aspects in the data collection process does appear useful, however.

7.1.3. L1-standard, variance, and text-linguistic effects

L1 is typically used as a gold standard in learner corpus studies. This has been criticized for over- and underuse studies (Gries and Adelman, 2014) with the argument that native speakers do not have a single modus operandi for all cases, and extended in Gries and Deshors (2014), in a mixed-effect model of the usage of gerund vs. infinitive complements in the International Corpus of English (ICE, Greenbaum (2014)). (Deshors and Gries, 2016) further extends the model to a comparison with ICLE, the International Corpus of Learner English, showing a very small difference in accuracy between the L1 model on L1 and the L1 model on L2 (goodness of fit = 0.81 L1 on L1, 0.76 L1 on L2, (Deshors and Gries, 2016, 201), accuracy = 88.5% L1, 85.2% L2). Based on 12 features from the syntactic, semantic and metadata domains (syntactic shape of the object, verb semantics, country of English variety, for example), two of which are open classes (lexeme of the VP, lexeme of the complement), the model allows for dozens of fixed variants multiplied by two open classes. Still, it does not perform extremely well. In the authors' own words:

> "The first analysis yielded a classification accuracy of 88.5%, *which is not much, but significantly* higher than the baselines of always choosing the more frequent complementation pattern (i.e., to) or choosing proportionally randomly (...). More illuminating is the analysis's C-value, which *just about* exceeds the usually-assumed threshold value for 'good' results of 0.8 with a value of 0.81", (Deshors and Gries, 2016, 201, my emphasis).

In other words, the model, in spite of accounting for dozens of factors, does not grasp the L1 phenomenon with impressive statistical accuracy. Moreover, it leaves out broad text-linguistic context like topic and register completely.

A number of metrics in chapters 4 and 6 have shown that native speakers in Kobalt, although they are from the exact same classroom and therefore peers in almost every respect,[15] show large differences in their language use. Some of this can be attributed to stylistic or register choice, like verbal vs. nominal style. In that sense, a learner who uses zero instances of categories like modal verbs or constructional verbs may be equally as 'native-like' as one for whom these make up 15% of all their verbs. Some of this variance can be attributed to a prompt response bias: If I choose to make my argument about the possibilities in today's world, I will likely use more modal verbs. This of course challenges the idea of a single L1-standard, but it still allows for the possibility of an L1-standard

[15] Like age, socio-economic status (which in Berlin is largely divided by neighborhood), language environment (urban, multilingual Berlin, although some may be from bilingual families while others are not), level of education, language input in school, etc.

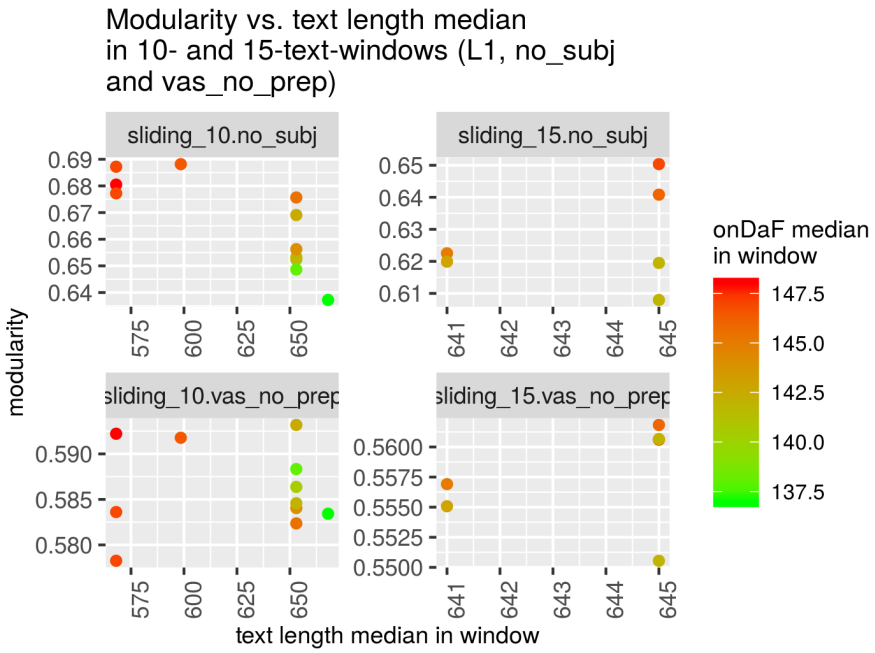


Figure 7.1.: Modularity by text length and onDaF median in L1, 10- (left) and 15-text-windows (right). Higher modularity for the no_subj-graphs exists for higher onDaF (red).

space, where while individual native speakers may use zero verbs in their constructional sense, statistically, they still use more than most learners.

However, some of this variance also likely stems from differences in language command. The participants were all 12th grade high school students, which is the most selective of the three to four branches of the German education system.[16] Those who have made it to 12th grade in high school are all at *Gymnasium* level, meaning that a year later they would reach *Abitur* level, qualifying them to study at any university in and outside of Europe. Yet still, some of those native speakers barely reach onDaF levels that are supposed to correspond to a C1-level in CEFR, a level some of their peers have nearly reached in their second language (English) by that time.[17] These are not isolated outliers, onDaF results are spread over a range of almost 20 points in Kobalt-L1. While cloze-test performance is an issue of practice to some degree, and learners may have been more motivated to show the full extent of their language skill in the test, this still leaves room for actual differences between native speakers. Sliding window analyses, which should be meaningless for L1, because onDaF is not supposed to measure actual differences in native speakers, show an effect of growing modularity in the no_subj condition for windows of 5, 10, and 15 texts (see fig. 7.1 for 10- and 15-text-windows).

If L1 exists on a spectrum, one of the questions for the analysis of learner language is how and where on that spectrum to locate a reasonable target space for SLA.

Chapter 2 referred to an existing discourse around *Bildungsdeutsch* or *Schuldeutsch*, 'ed-

[16] The other ones are *Hauptschule* and *Realschule*, which historically used to prepare adolescents to pursue vocational training in blue-collar trades and service industries, and where it still exists, *Förderschule*, for children and adoloscents with learning disabilities and special needs. *Gesamtschulen*, comprehensive schools, have mixed classrooms, but students remaining in school after year 10 are at *Gymnasium* level.

[17] At least this is a prerequisite for many English Studies programs in German universities.

ucated or academic German' and 'school German' and the dependency of success in the German education system on the command of this register (Petersen, 2014; Haberzettl, 2016; Cantone and Haberzettl, 2009). This is (unfairly) discussed mostly as a problematic area for bilingual students in Germany and hence mostly studied in bilinguals (Müller et al., 2016; Haberzettl, 2009; Schulz et al., 2008).[18] Hence there is little to no research into L1 variation in lexicosyntax in adolescent or young adult speakers of German who are not considered *Bildungsverlierer* ('losers of the education system'), and that is not dialectal or otherwise determined by variational strata.

One side aspect with the issue of variation in L1 and L2 is that register choice itself may be a dependent variable underlying the same avoidance mechanisms as other linguistic factors: A writer who feels insecure in one register might choose a different one that feels more secure. For example, a writer who feels linguistically secure listing possibilities, but not arguing on a more abstract level, will choose to do so, even if they may be able to think of more compelling arguments in response to the prompt. Perhaps Belarusian learners feel *linguistically* insecure writing in their L2 about more abstract spheres such as society and politics until relatively late in their language development, and default to listing opportunities of people in the past vs. now. Importantly, their sense of insecurity may or may not correspond to their syntactic or lexicosyntactic abilities. Perhaps Chinese learners receive more training in describing aspects of their country or evaluating abstract statements in their language classes, and therefore feel more secure with it and default to this. Then it might not be register *choices* as much as register or topic *constraints* that shape the form of the final text. The same may be true of the native speakers, although native speakers may have more islands that they feel comfortable to choose from, overall creating a more evaluative or analytical impression.

This has greater implications:
In chapter 4, where ΔP values for OBJP slots were discussed, native speakers showed more combinations that are clearly lexicalized in German (like *zur Verfügung stehen*, 'be at the disposal of') compared to learners, but still at small numbers. How could we tell that these are indeed reflective of register choice, and not lexical or lexicosyntactic teddy bears, similarly to BEL learners' *eine Ausbildung bekommen* ('to get an education', which is unidiomatic in German)? How to tell if a register is a register, i.e. a holistic and structural property of text, and does not only partially mimick one?

So far, the presence of register differences has been derived from diverging lexical distributions between L1, CH, and BEL. However, it is also possible that *none* of those groups are actually competent in the presumed register, if register is understood not merely as a surface lexicosyntactic category (choosing the right words in the right constructions), but a larger, text-linguistic unit that includes structures of higher abstraction levels, too (such as anaphoricity, referentiality, discourse and information structure, lexical vs. implicit cohesion, etc.). Even if three registers (more narrative, more descriptive, more evaluative) can be shown to exist in Kobalt on a larger scale, register may still differ not only between texts, but also in text parts. This is also what was suggested in chapter 6 in the VAS sampling, where text structural effects were found systematically for learners. Effects in L1 are less systematic, but still present.

Register may also not necessarily be *chosen*, but could be *entailed* by the argument, or

[18] Petersen (2014) and Haberzettl (2016) do in fact compare bilinguals with monolinguals and conclude that main differences can be explained from age and from socio-economic status and education level of the parents rather than mono-/bilingualism of the students.

even a result of self-priming. This would then also contribute to the explanation of register shifts as they have been shown to occur in BEL. Even a slight shift in the argumentation might prime a different frame altogether. A similar idea has been proposed by Hoey (2004, 174):

> "What I want to claim in this paper is that every lexical item is primed for use in textual organisation. The notion of priming is taken from psychology and in this context means that our encounters with a word accustom us to expect it to be used in certain kinds of ways to such an extent that these potential uses become part of our knowledge of the word and to some extent constrain the way we are likely to use the word ourselves.
>
> More specifically I want to make the following claims:
>
> 1. Every lexical item (or combination of lexical items) may have a positive or negative preference for participating in cohesive chains.
> 2. Every lexical item (or combination of lexical items) may have a positive or negative preference for occurring as a part of Theme in a Theme-Rheme relation.
> 3. Every lexical item (or combination of lexical items) may have a positive or negative preference for occurring as part of a specific type of semantic relation, e.g. contrast, time sequence, exemplification.
> 4. Every lexical item (or combination of lexical items) may have a positive or negative preference for occurring at the beginning or end of an independently recognised 'chunk' of text, e.g. the paragraph.
> 5. If a lexical item (or combination of lexical items) has any of the above preferences, it may only or especially be operative in texts of a particular type of genre or designed for a particular community of users, e.g. academic papers."

In this reasoning, the choice of a word primes a context that comes with specifications of a number of linguistic levels, or in Hoey's words: "This is not to say that the choice of a lexical item compels certain textual developments but it certainly makes those developments more likely" (Hoey, 2004, 189).

Similarly to the conceptual shift from an independent view of syntactic and lexical modules to the idea of lexicogrammar as interdependent, this suggests an integrated approach to lexicsyntax and text linguistics: Perhaps text is not ideally seen as a process of text generation that is filled with lexical and syntactic material, but rather, lexical and syntactic material shape the text over the course of its production, leading to more similarities in learner groups where, for lack of words, more similar contexts are primed.

If shown systematically on fresh data, this would provide direct evidence to the concept of non-ergodicity in natural language (the possibility of getting caught in a particular subsystem of language from which relative frequencies can no longer converge to overall limits or probabilities). It would also raise new questions towards the dynamics of text, burstiness, self-priming, and require a discussion of when and to which degree register "choice" in learners could be considered an epiphenomenon of lexicosyntactic command.

The linguistic understanding of first language acquisition has begun from an adult perspective, seeing toddler's language production as a kind of adult language production with gaps, and has moved towards a child-centric description, seeing toddler's language production as something that is built from similar blocks as adult language, but does liekly

not follow the same abstractions and conceptual structures. Perhaps it is helpful to look at learner text production in the same way, namely as reminiscent of a certain register without suggesting it is *already* functionally similar to the L1 equivalent.[19]

All of these remarks are observational and post-hoc. In order to learn more about the dynamics, a clearer model of register in learner language is required, along with a more specified model of register as a product of a self-constructing process, and annotations thereof in data that has not been multiply tested. Ideally, that data would also include language production in different (attempted) registers, which would help to identify lexical or lexicosyntactic preferences over actual register differences. In that sense, comparing high school students and not too advanced L2-learners might in fact provide comparability, because both lack command of the target register as it is defined in higher levels of German (for example in academic studies).

7.2. Methodology

This section outlines paths for future research in the methodological respect of this study. It first discusses the necessity for replication and extension of modularity-based analyses to new and differently shaped data, and presents an overview of desirable improvements to the underlying linguistic model in future research. It concludes with a more general point pertaining to the theory of modeling and methodology in corpus linguistics in discussing the issue of data size and the tension between deep linguistic analysis vs. the compilation of large corpora.

7.2.1. Replication

The first step in establishing the method suggested in this thesis lies in replication and extension. Replication is generally necessary in empirical research.[20] Here, it is of particular importance for two reasons:

1. A new method was introduced and internally validated. But the consistency of a dataset in a specific regard does not validate results externally, i.e. relative to a group or linguistic aspect per se. It is still possible that the dataset was idiosyncratic. This means that in order to establish Louvain modularity as a usable method in corpus linguistics, it must first be validated externally.

2. While the hypotheses were derived independently of the data, the method was developed from the insight that an item-based approach would not be helpful with the size and the dispersion of the data at hand. It is therefore possible that, on some abstract level, a confirmation bias has found its way into the study. This can only be ruled out through replication and extension to unseen data. In a first step, data should be extended to include Russian monolingual data for a comparison with the Belarusian learners; and to onDaF-earlier Chinese learner for a clarification of the trajectory.

Moreover and specifically, an extension to other topics, registers, and corpus sizes in both L1 and L2 is necessary not only to verify the acceptability of the metric, but also to gain a

[19] No similarity between children and learners in SLA implied.
[20] See Plonsky (2014) for a discussion of linguistic methodology and replication

better understanding of its mechanics. What does a difference of 0.05 in modularity truly mean? How does modularity play out in larger corpora, where more of the distribution is more evenly filled – unlike in small corpora, where the very frequent lexemes still appear frequently, but most of the rest of the distribution is filled with hapaxes, and intermediate frequencies are nearly non-existent?. Does modularity in larger corpora indeed converge after a limited number of texts, as it has been suggested in the previous chapter, or was that a misreading of the data? Does it perhaps even approximate zero? If it converges, then when and at which value, and how does this differ between text types, registers, L1/L2 combinations, and so on? Are there specific modularity values inherent to text types, registers, languages, etc.?

This thesis aims at contributing to the development of a framework of methodological development, i.e. to the establishment of a critical discourse around epistemological aspects of methodology. It strives, eventually, to contribute to a quantitative corpus methodology that is validated, epistemologically sound, and as closely in line with the linguistic model as possible, i.e. a methodology at the quantitative-qualitative interface. Methods are not merely toolboxes, but definitions of the interface between a theoretical and an empirical understanding of linguistic data. As such, they are central to the development of better models, and should be carefully chosen and validated. In this study, a data-internal validation against frequent confounding factors like text length, corpus size, and grouping marks a first step, but an external validation of all of these aspects, and a verification of the usefulness of the method in the first place, is due.

7.2.2. Improvements to the linguistic model

The linguistic model underlying the graphs as it was developed in chapter 5 is simplified in many ways. Future extensions of this research should aspire to account for the following aspects and

- look into the role of pronouns and their role in subject and object slots;
- more clearly separate semantic subjects through the inclusion of the unergative/ unaccusative distinction (Kuno and Takami, 2004);
- include word sense disambiguation for a comparison with a non-disambiguated graph. This would be interesting with respect to the role of homonymy or homo-graphy in entrenchment;
- account for the semantic specificity of verbs in the analysis (complex vs. simplex, for example, but also constructions);
- consider categories of complex predicates (Müller, 2002) and constructions;
- account for subjects in coordinated claues and other structures that do not ideally represent subject dependency in dependency grammar. This should be used to further clarify the differences between the no_subj and the vas_no_prep graphs and some inconsistencies in the interpretation that have been mentioned.

Some more complex aspects have also been mentioned as usefully extendable for a deeper understanding of coselection:

- A classification and systematization of different types of coselections in a multidimensional model, including a disentanglement of different properties such as convention,

idiosyncrasy, non-compositionality, flexibility, frequency of occurrence, salience, prototypicality, etc.;

- a differentiation between conventionalized and/or preferred vs. constrained coselections including their morphophonotactic aspects, and a clarification of the relationship between productivity and coselectional constraint;
- the role of emergent vs. individual effects, particularly where graphs are more modular in smaller windows, but only in some of the onDaF ranges;
- differences in processes at very early intermediate/upper-beginning stages, where modularity was *much* lower than shortly after and the relationship between a more item-based vs. a more structural development and how it would become apparent in terms of modularity measurement.

Unlike the previous points, these require both new data and an extension of existing linguistic models, i.e. further integration of different strands of usage-based linguistics, phraseology, and general SLA, as well as more methodological and quantitative modeling and conceptual validation.

7.2.3. Larger data and sampling

The validity of the concerns raised about the statistical analysis of lexicosyntax, the validity of the graph-based analysis itself, as well as the interpretation of the findings around variance and group vs. individual effects hinges on one central question: Would a larger dataset have been suited to solve the major limitations of the study? Or in other words: Was Kobalt too small for the analysis attempted in this study? This is an important question with respect to corpus linguistics in general, since all methodological planning depends on the availability of resources, and clarity about the threshold of data quality and quantity for any specific method is crucial for its successful employment.

Let it first be said that, while the subcorpora in Kobalt divided by language and onDaF criteria are indeed rather small, Kobalt as a whole is not much smaller than other corpora used in learner corpus research: Granger and Bestgen (2017) use a corpus of 223 texts extracted from ICLE (International Corpus of Learner English, Granger et al. (2009)), where texts are of 500 to 900 tokens in length and divided into three L1 groups (74 French, 71 German, 78 Spanish). Minus one language group, that is about as large as the Kobalt subcorpora, even a little smaller than the BEL subcorpus of Kobalt (BEL=89, CH=62). Römer et al. (2014) for their study of verb fillers in verb-argument constructions used larger written corpora (236 and 198 thousand tokens depending on the L1 of the learners), but the spoken corpora were less than twice the size of Kobalt (63 and 86 thousand tokens vs. about 36 and 56 thousand in Kobalt). Paquot (2013) in her study of L1-transfer of lexical bundles in ICLE essays written by learners with L1 French uses a sample of 228 essays at a mean text length of 598 token (p. 397). Since she rules out topic-related bundles, she arrives at top frequencies of 22 occurrences of a bundle, which is for the very general *we can say* (p. 402). This is not very different from Kobalt. Gries and Wulff (2009) in their study of verb + gerund vs. verb + infinitive alternations work with 480 instances of the gerund and 2863 instances of infinitive complements. But then these numbers relate to only two categories. Divided by verbs, they end up with 48 and 98 verb types respectively, of which the large majority will be hapaxes and unique coselections. In addition, German

is also studied much less than English in the world, and to my knowledge, no larger corpus of learner German that is as homogeneous and as controlled as Kobalt has been compiled to date.[21]

But aside from pragmatic choice and common practices, the question is still worth discussing: Would an analysis of coselectional preferences profit massively from larger data?

It was mentioned in chapter 4 that there are epistemological problems with lexical association measures because it is unclear whether language is a stationary and ergodic system, i.e. a system in which relative frequencies approximate expected values and thus map to stable probabilities, and where it is impossible to get caught in an idiosyncratic corner or bubble in which relative frequencies will not ever approximate cross-system limits.

If this were merely a philosophical issue concerning changes at a grand scale of billions of tokens, one might wish to accept an approximation through large corpora. But the problem is that even in smaller corpora like Kobalt, this quickly becomes relevant even within the scope of a study like this. No-one would likely argue, even if lexemes had persistent probabilities, that those would be reached in a smallish corpus like Kobalt. But what kind of text could be used to extend Kobalt in a way that yields reliable convergence of lexeme frequency? Another German learner corpus, Falko (Reznicek et al., 2010), is larger and served as a template for Kobalt. It containts texts written in response to five different prompts touching on controversial topics, like whether people should be paid in accordance with what they have contributed to society, whether feminism has done more harm than good, and whether criminal activity pays off or not. Essays on those topics obviously lead to five different lexical distributions, meaning that in a combination of Falko and Kobalt, lexeme frequencies as measured in Kobalt would decidedly *not* be brought closer to convergence, but actually drop for most of the more frequent lexemes. Aside from rather generic and functional words, the same is probably true of any extension – even if Kobalt was extended with the same prompt, but now, seven to eight years after the initial data collection, topics would likely change drastically, because the world has changed, too.

This of course is the question of whether corpora are representative samples of natural language, which has been discussed most prominently in Biber (1993), Kilgarriff (2005), and in Evert's libary metaphor (Evert, 2006). More recently, Koplenig (2017) criticized this with respect to whether a corpus can ever be representative of language distribution per se, and suggests that since it is impossible to tell what distributions would look like in a "comprehensive library of a language", it is impossible to map to those distributions in statistical tests. I would go a step further and say: Kobalt is not only not a random sample, it is not a sample at all: A sample should not change the total of the subject.

Koplenig uses the example of red and blue flowers with four and five petals, discussing conditions of their independence if one collected a number of red and blue flowers in the wild. Growing these these flowers in the lab, numbers might change, but this could likely

[21] There are several more German learner corpora: KanDel, the Kansas Developmental Learner Corpus, compiled from writing samples from beginning learners of German at undergraduate level with different prompts (Vyatkina, 2016); Falko, Fehlerannotiertes Lernerkorpus, compiled from very advanced learners of German at undergraduate level, also in response to different prompts (Reznicek et al., 2010); and Merlin, a corpus of L2 compiled from response texts to a standardized test, (Boyd et al., 2014). Out of these, Merlin is the largest, with over a thousand texts for German L2, but even this is barely a magnitude larger than Kobalt, topic and register are not controlled for, and some of the texts are rather short because they are produced by A1 and A2 learners.

still be accounted for. But what if in my lab, I intentionally or unintentionally grow six-petaled flowers, or flowers of other colors? No statistical test could infer from those to an outwardly existing population of four- or five-petaled flowers. The same is true of quasi-experimental data as it is used in many learner corpus studies: There would not have been 151 learner texts on the matter of whether previous generations had a better life than today's youth in a similar context if it had not been for the Kobalt project collecting them. This is not necessarily true of most L1 corpora (like newspaper or historical corpora), but for the domain of learner corpora it often is. This means that relative frequencies in German learner language as it is documented are not only not stable, but they can be actively changed through the act of prompting more learners to write texts on that topic. Most learner language that exists outside of such writing would be conversational, work-related, academic, or of other types, and lexical frequencies would be very different. But if we were to extend Kobalt ad infinitium, the result would be a significant change of German learner language *as it is realized* or documented. This would constitute a non-ergodic bubble in which lexical frequencies substantially deviate from all other contexts. A large Kobalt based on the same prompt would then *be less representative* of learner language as it exists.

Often, corpus data is collected from naturalistic usage and not created for the corpus. But the same process that is performed by the Kobalt project is also initiated by other developments and discourses in a population of speakers: If there is a demand for certain kinds of novels, people will write them, thereby changing the absolute frequencies of some, but the relative frequencies of *all* words in the language. Even on a small scale, this means that extension to larger data can be ontologically and epistemologically challenging.

Unlike what it may seem, this is actually an optimistic observation, because it alleviates the pressure of creating representativity through balancing large corpora. Large corpora are resource-intensive in compilation, annotation and hosting. Many interesting aspects of linguistics cannot ever be annotated in large corpora, such as rhetorical structures, phonetic, or pragmatic aspects. These are either very fine-grained and require enourmous transcription effort, like phonetic research often does, or very intertwined and ambiguous, and require a lot of hermeneutic negotiation, like pragmatic and text-linguistic annotations often do. A smallish corpus like Kobalt can be read and manually corrected in its entirety, but already at less than one higher magnitude, at perhaps 500 texts, this would be impossible to accomplish in a doctoral thesis. This also means that smaller corpora allow for the consideration of more factors in the data collection, and thus for deeper linguistic analysis, since more noise can be separated from the signal in collection. Large corpora, on the other hand, are structurally unable to exhaust the full potential of linguistic analysis.

Obviously, some corpora are still too small for a quantiative analysis. No matter how good my model, a quantiative analysis of the total of 12 lines in the two *Merseburger Zaubersprüche* (Merseburg charms)[22] will not yield fascinating insights. However, it seems plausible from this study and others[23] that there exists a middle ground between corpora that are indeed to small for quantative analyses and very large corpora that require lin-

[22]Charms written in Old High German that were found in the library of Merseburg and date back to the 800s or 900s, for an overview of newer publications, see Düwel and Heizmann (2009).
[23]For example (Hirschmann, 2015; Wan, in prep.)

guistic concessions in compilation and analysis. This middle ground can be conquered by developing specified, formal, and quantitative models and methods and integrating them with qualitative and hermeneutic approaches. For this to be successful, an epistemological discussion addressing the question of appropriate application areas for nomothetic vs. idiothetic explanations is required, a distinction descriptive of the scope of an analysis (cross-systematic, generalizable vs. detailed, but related to only a subsystem). Qualitative and quantiative research is often described as a continuum (Newman et al., 1998; Niglas, 2007; Onwuegbuzie and Leech, 2005), and the very center of this continuum, where the two concepts meet, might be intrinsically well-suited for the study of corpus linguistics.

With this said, there are two aspects in which a study like this would indeed profit from more data: First, it would be helpful to test the method developed in this thesis on data of different sizes to gain a better understanding of the mechanics, which is a matter of methodological validation. Secondly, linguistically, more data would be helpful for assessing the potentials and limits of an exemplar- or item-based analysis of coselection. This quickly becomes problematic, because for coselections that occur only once or twice in the ten to twenty texts per subcorpus, a corpus extension limited to another few hundred texts would likely not fill up the distribution in the desired way. It would not necessarily raise frequencies of the already observed coselections, but add many more hapaxes, which means that for a more comprehensive understanding of exemplars in coselection an extension, extension by at least one or two magnitudes is necessary. Yet, this would result in losing control over annotation precision, and to the bizarre case of skewing the existing and documented writing of German SLA to the overwhelming case of a single type of text and a single topic. This creates a non-ergodic bubble as described above (or, to stick with an earlier metaphor, a very narrow-ranged zoo). But even then, most of the newly found coselections would still only occur in a small minority of texts. This means that individual effects would likely grow strong against group effects, carrying more statistical problems. One way around this is to study exemplar-based coselection in use in true longitudinal data, but this on the other hand entails the problem of topic-specific lexical distributions, too (since participants writing on the same topic several times would likely not produce the same texts as they would have at the same time without the previous writing experience). It also carries the problems of collecting rich longitudinal data in the first place (keeping participants engaged, providing resources over year, (dis-)continuity of progress, etc.).

This is a fascinating problem in light of the fact that conventional coselection has made a career from the periphery of language description in generative approaches to becoming one of its centrally observed aspects in usage-based approaches to lexicosyntax or in the idiom principle – yet still, its measurement in exemplars remains a complex and challenging task.

7.3. Graph-based modeling of linguistic phenomena

As outlined in chapter 5, graph-based modeling is not common in linguistic research at present. Graphs model entities and their relations at maximum abstraction, which explains their wide employment in a range of quantitative fields. At the same time, maximum abstraction is not always desirable in linguistics, since many phenomena are better understood as an interaction of more abstract and more concrete, item-specific workings. This section first reviews this issue in the context of coselection, where isographs – graphs

with identical structure but different content – may pose a challenge to the correct identification of useful levels of analysis. It then presents ways for further inclusion of graph-based modeling and network analysis in linguistic research. This includes a discussion of graphs as a potential solution to the challenges of non-ergodicity, the employment of graphs as an alternative to other quantitative methods where data is sparse, and a sketch of grammar as graph that allows for the explicit modeling of grammar and lexis within the same space as well as an understanding of association strength as local (rather than cross-system) probabilities.

7.3.1. The isograph problem

The analysis presented in this thesis cannot conclusively verify concept validity: It is unclear whether there is a perfect mapping between the linguistic features contributing to coselectional constraint and its quantification through graph modularity. Part of this is due to the lack of a linguistic understanding of coselection beyond the concept of strength of association between elements. A multidimensional model may come to the conclusion that syntactic, text-linguistic, and phonotactic features ought to be considered in the model more strongly. The analysis performed in this work is likely undercomplex with respect to those aspects.

A different problem lies in the abstraction of graphs as such: The measurement of modularity has not technically shown a development of coselectional constraint in a certain way in Kobalt. It has only shown different measures of graph structure as defined in the model of lexicosyntactic coselection in chapter 5.

Why are those not the same? In a graph as it is accepted into the modularity computation, all nodes are equal. The metric has no way of telling frequent from infrequent or productive from unproductive lexemes or structures, aside from the degree of a node, it looks only at the structure itself. The graph-theoretical concept of isographs describes (sub-)graphs that can be mapped onto one another, i.e. that are structurally identical. Since graph theory is not concerned with properties of individual nodes, only classes of nodes, any item-specific information is lost in this comparison. For example, if we took a set of a verb and seven arguments, that subgraph would be the same in the computation and in graph theory in general, regardless of whether that verb was highly frequent and unselective like *haben* ('to have') or infrequent and highly selective in other corpora – phraseologically, these cases are rather different.

Louvain modularity as a metric of lexicosyntactic development works in accordance with the hypotheses in Kobalt at least for the BEL learners, but this does not necessarily mean it maps to the same concepts. A proof of concept may lie in the observation of a general process of randomization, diversification and specialization between early intermediate and advanced stages of learner German as presented in chapter 4, which is captured in essence by the computations of Louvain modularity in chapter 6. However, the whole phraseological extent of the *idiom principle* (Sinclair, 1991) could not be observed in Kobalt and it is plausible to assume that coselectional constraint may not map *in full* to Louvain modularity, or in other words, that modularity measures only parts of it.

This needs to be validated in future research, ideally in a larger and more homogeneous corpus that allows for both an item-based and a distributional view of the data. Specified hypotheses for modularity as a measure of generalized processes, for distributional coselectional constraint, and for item-based coselectional constraint would have to be developed to validate for these factors.

An interesting question in this respect is whether L1-like subgraphs can be identified in L2 corpora, both bound to concrete items and structurally without item specification; if L2 graphs can be decomposed into target-like subgraphs and transferred subgraphs from the L1 of the learners; and whether groupwise variation between learners could be predicted from such subgraph merges. It would also be interesting to see if subgraphs differ by semantic verb group, such as complex verbs vs. simplex verbs, as suggested by (Plank, 1984), and whether processes of specialization can be traced through isomorphism analysis.

While the detection of subgraph isomorphisms is NP-hard, several algorithms have been proposed in recent years that solve the problem somewhat efficiently (Emmert-Streib et al., 2016; Rivero and Jamil, 2017; McKay and Piperno, 2014) and quantifications of similarity can also be applied through graph similarity (graph edit distance, Bai et al. (2018); Fischer et al. (2015); D based on connectivity and information flow, Schieber et al. (2017), and distance based on the maximal common subgraph Bunke and Shearer (1998), among others). Also, while the structural problem with unspecified subgraphs is computationally hard, an item-based comparison is rather trivial, requiring only a matching of subgraphs, i.e. edges.

Modeling a corpus as a graph is not hard as long as it is well parsed[24] and downloadable. While computations are somewhat expensive, an analysis of this kind would still require limited resources and might provide interesting insight into structural aspects of lexicosyntactic coselection from many angles.

7.3.2. Graph theory and network analysis in linguistics

Graph theory is essentially absent from present day linguistics, and graphs and network analysis are barely recognized as modeling and analytical tools by linguists beyond their employment in the visualization of analysis or data exploration. Almost bizarre against this background, the following quote from 1961 suggests their potential once used to receive more attention Goodman (1961, 55):

> "If it seems at the present stage of structural linguistics that nothing more will ever be needed than the familiar rudiments of graph-theory, it probably seemed at a comparable stage in the development of physics that nothing more would ever be needed than elementary arithmetic".

Where graphs are used in language modeling, it seems that fewer linguists than psychologists are involved (although there is some overlap with psycholinguists, see for example Ellis et al. (2014); Beckage et al. (2010); Kenett et al. (2016); Chan and Vitevitch (2010); De Deyne et al. (2016)), as well as physicists (Ke, 2007; Martinčić-Ipšić et al., 2016; Cong and Liu, 2014; Wachs-Lopes and Rodrigues, 2016), or mathematicians (Mehler, 2008; Mehler et al., 2016). This might be why where network analysis is used on language, it typically relies on the same few and limited measures like degree distribution and clustering coefficients (Wachs-Lopes and Rodrigues (2016); Ferrer i Cancho and Solé (2001); Ferrer i Cancho et al. (2004); Li et al. (2005) and Choudhury and Mukherjee (2009) with references to many more studies into the same structural properties), and is limited to modeling words and word co-occurrences in positional or somewhat basic syntactic ways rather than integrating more linguistic depth. Degree distribution and with it clustering coefficients are themselves limited by the long-tailed distribution of lexemes – a hapax can

[24] This of course often presents a challenge to available resources.

only have a degree of one, whereas a frequent word will necessarily have a high degree, meaning that degree distribution is a function of the frequency and the positional context included in the analysis: an epiphenomenon.

However, holding a potential far beyond this kind of application, graph theory and network analysis are interesting to look into anywhere that there is a relationship-centered problem; and models based on graphs and graph metrics may capture developments in ways that are more in line with linguistic theory than statistical approaches often are. In fact, the opposite of an ergodic system is a system that evolves and changes constantly from factors within and without, a *complex dynamic system*. In a complex dynamic system, and in system theory in general, randomness is not assumed as a constituting part of the system (which is not the same as not allowing for any randomness at all). Rather, a system at its very simplest is defined by a number of items $X_1 - X_i$ and *a number of relations* $R_1 - R_n$ *between those items* as defined in general systems theory (Mesarovic, 1964, 6-7). This is a mathematical description of the Aristotelian truth that the whole is more than the sum of its parts. A group of items and their relations *is* a graph.

The challenge for successful modeling in this framework lies in the identification of all relevant relations between the items, some of which may be superpositioned and emerge from the interplay of the others. There have been a number of papers authored by well-known linguists stating their view of language as a complex adaptive (= dynamic) system (Steels, 2000; Mislevy and Yin, 2009; Massip-Bonet, 2013; Holland et al., 2005, among others). Most notably, this idea has been formulated in a position paper by the 'Five Graces Group' involving Joan Bybee, Nick Ellis, Diane Larsen-Freeman, William H. Croft and others (Five Graces Group et al., 2009) – a group quite representative of usage-based linguistics and language variation studies at the time. Yet, there has been very little discussion of the repercussions of this view on the methodological turn towards more modeling and quantitative analysis that has been ongoing in linguistics for the past twenty-odd years. Rather, the idea of the complex adaptive system has been suggested as a positional statement that *turns away* from explicit modeling by allowing for a large number of complexly interwoven subprocesses that cannot always be predicted. But

> "this is the problem of overestimating the flexibility of the system and thereby failing to recognize the existence of its significant regularities (...). That is, on the one hand, dynamical systems may at times behave in ways that are difficult to predict due to the complex nature of the interconnectedness of their components, yet, on the other hand, they are not infinitely open ended and flexible. Moreover, as Meara (personal communication, 2009) has pointed out, sometimes behavior seems to be complex, yet the apparent complexity may, in fact, be understood in terms of the operation of simple elements, removing the need to appeal to an underlying complex system" (Segalowitz, 2010, 19).

Thus, the question of ergodicity in language is not limited to learning whether a given subset of language is ergodic or not, or to the philosophical and epistemological dimensions outlined in previous sections. Rather, it holds the potential for a fascinating future for usage-based, well-modeled quantitative and/or exact linguistics. If a system is partially ergodic, it can in some cases be modeled to oscillate between different states that *on average* do in fact reach expected values, as has been shown for opinion dynamics in social networks by Frasca et al. (2013); or to contain ergodic and non-ergodic subsystems that coexist and coevolve in specific ways, as has been shown in molecular processes in the plasma membrane in biology (Weigel et al., 2011). Showing either for specified

subspaces of language would allow for statistical comparison in a well-defined mathemtical space, providing a much needed foundation for the validity of those approaches in corpus linguistics.

Looking into stationarity and ergodicity from a linguistic view can bring a new lense into linguistics through which language development and even synchronic dynamics can be much more intricately, but still formally and quantitatively modeled as a synthesis of several subsystems or subfields. This would constitute a step forward from a methodology that requires averages over large samples, which often enforce a broad and rather unlinguistic categorization or the acceptance of assumptions that clearly contradict linguistic analyses. Of course this would also require the development of language-specific methods that are able to adequately capture those phenomena. Graphs might provide a framework for a formal model of such spaces and the dynamics therein.

This is not meant to imply that graphs are by definition always a good choice. In fact, the recent enthusiasm for graphs in the field of digital humanities does not seem to incorporate a lot of formal modeling. In the spirit of "all models are wrong, but some are useful", the question that needs to be asked is: What constitutes – from a *linguistic* point of view – a good quantification and operationalization of concepts in question? This is far from trivial: There is a plethora of lexical association measures that are used in much of linguistic research, somewhat helplessly at times, as is discussed by Gries (2019) – and his most recent proposal to combine them in tuples does not appear much closer to an actual solution either. And there is a lot of surface-oriented network analysis of language, mostly coming from other fields and with very little linguistic modeling involved, which does not seem to provide much insight into the dynamics of language specifically. The problem then might not be in the methods per se, but in the conceptual mappings between theory, research question, and the demands of the mathematical model. Maybe it is just not a very insightful approach to average over large, but heterogeneous data, because the underlying system works in a different way, and one that is in fact already better understood by linguistics than is represented in its operationalizations. Graphs might provide such a way for core-linguistic modeling that remains both formal and linguistic. One example for a usage-based, graph-based model useful in linguistic analysis will be outlined in the next subsection.

7.3.3. Grammar as graph

The model employed in this study is primarily lexical and only implies aspects of syntax through the edges defined as possible by the filtering function: In the Kobalt graphs that were used to compute modularity, nouns cannot be connected, for example, and subgraphs based on edge labels are realizations of rules of grammar or the output of a function of grammar rules. It is, however, also feasible to model syntax more explicitly:

For one thing, subgraphs divided by constructions or argument slots can be considered separately. It was hypothesized and partially shown in chapter 4 that different argument slots exhibit different degrees of constraint. Of course it is also possible to look only at the coselections of a single slot with all its verbs and arguments. In Kobalt, this quickly leaves tiny graphs (which are naturally more modular than large graphs) for the less frequent slots, which is why in the analysis here all argument slots were considered at once. An analysis of individual argument slots in Kobalt would, however, allow for a more qualitative evaluation.

Secondly, argument slots or constructional senses may be modeled as nodes of their own,

where lexemes can either connect to both the syntax nodes and the argument lexemes and their nodes, or only to the grammatical categories. This is illustrated in fig. 7.2.

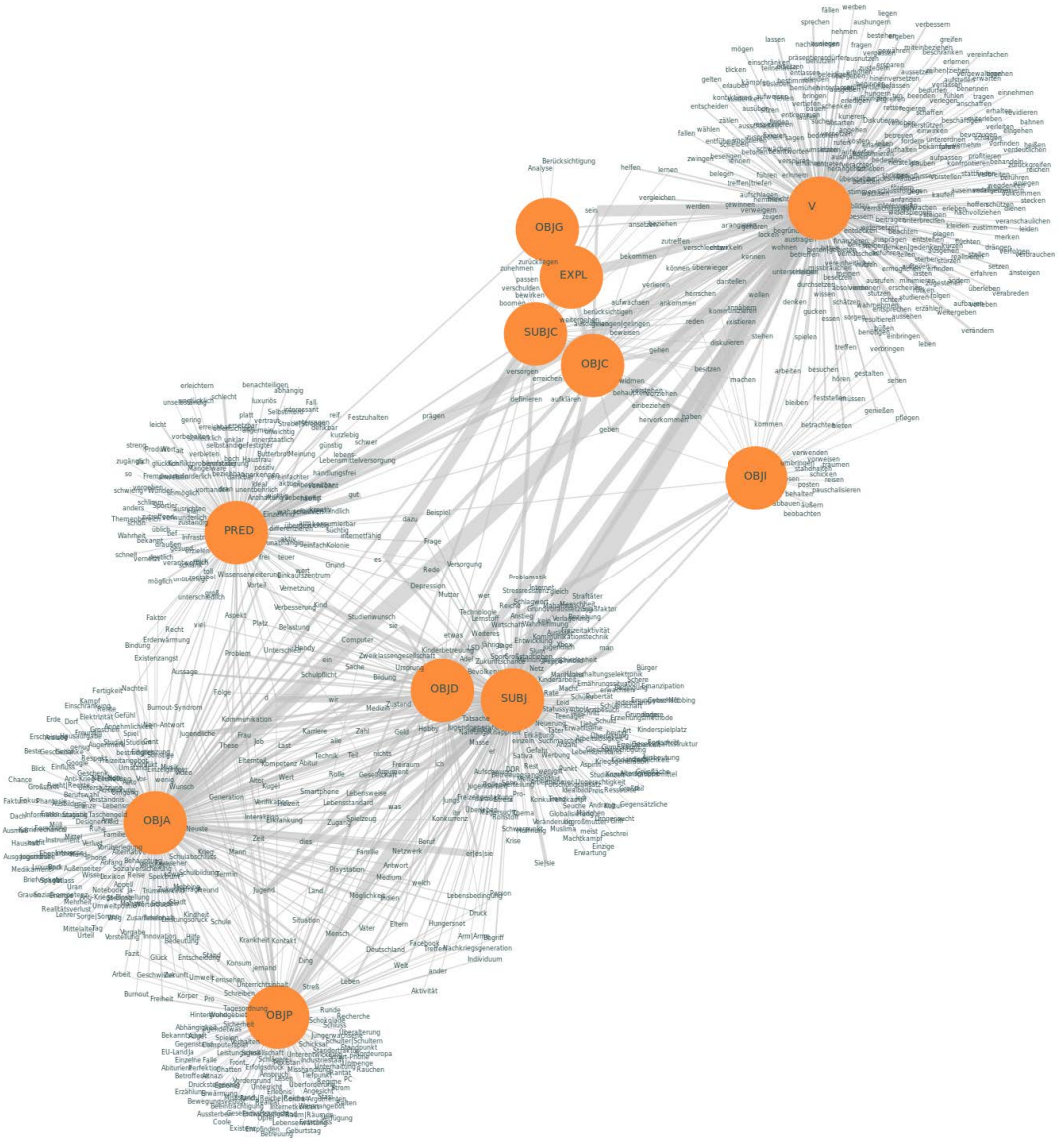


Figure 7.2.: A graph model of lexicosyntax in Kobalt L1 (based on the vas_no_prep graph). If slots share more lexemes or if they share lexemes that occur frequently in both slots, they are drawn closer together. Otherwise they are torn apart by the pull of the lexemes located elsewhere.

A graph-based grammar model also allows for the explicit modeling of intermittent construction levels, so instead of abstracting to OBJA, to model distransitive structures as an intermediate level, or even smaller, specified constructions. These could then be viewed as an instance of OBJA instead of or in addition to the lexemes, depending on the research question. This approach is very similar to what Zeldes (2012, chapter 6) models as the cognitive model of productivity. The claim here, however, is not cognitive. While it may

also prove true that a graph model of grammar is a good representation of cognitive structures,[25] the idea outlined refers to a model of grammar in use: Of course, conceptually, it is not new at all, in fact, it is already theoretically formulated in the shape of inheritance networks in construction grammar (Goldberg, 2006; Lasch and Ziem, 2014; Zeldes, 2013a) and in emergentist models of lexicosyntax (Hoey, 2012; Ellis and Ferreira-Junior, 2009; Ellis, 1996; Tomasello, 2009). Graphs are also generally very inviting as a metaphor because they model relationships, and any systematic perspective aims to also model relationships. But what is meant here is not a *metaphorical*, but a *formal* formulation of grammar as graph, i.e. the exact representation of grammar as it occurs in a corpus in a graph model, and hypothesis-based, *quantitative* research on those.

In a grammar-as-graph model, modularity is unlikely to be of great use, because all nodes are fairly interconnected through the large hubs of the grammatical categories. However, what can be made visible are the forces that lexemes exhibit on the grammatical categories. This can be modeled in any detail, morphosyntactically, constructionally, coselectionally, etc., simply through a connection of categories with instances. It is also possible to view these as a time series over different states of the target language or different historical stages of a language, for example. It has been noted that, while usage-based linguistics in general does not assume a discontinuity between lexicon and syntax, practically all models do model such a discontinuity in assuming constructions of different sizes which may or may not act as words. A grammar-as-graph approach could provide a solution to that by defining grammar and also words as subgraphs (words would be subgraphs of syllables or phones), or by explicitly modeling each presumed level and link those with edges.

A grammar-as-graph model also allows for a perk that appears intuitively useful but has not yet been formally modeled to my knowledge: Through edge weights, the most frequent collocates of a word or a structure can be modeled structurally. I would assume that any verb has between maybe 3 and 10 structurally encoded and therefore accessible collocates, while all other collocates are subsumed under a productivity category. Productivity and variance in coselectional density tend to mess with stationary probabilities across the whole lexical system or corpus. But here, distributional aspects would not have to converge cross-systemically. Instead, they can be used as *local* probabilities.They need not add up to one across a system, meaning local probabilities can also change and new words can be introduced without shaking up all probabilities of a system. My intuition is that this is also closer to what statistical models already try to recreate in linguistics: Local distributions that are shaped and influenced mostly through their neighboring structures rather than all other words related or unrelated.

Finally, coselection in grammar-as-graph could would not be limited to being modeled as a coordinated selection of two items in one simultaneous action of lexical retrieval, but could be seen as the semantically and morpho-phonotactically guided selection of *one* and the *entailment of the other* as a *default companion* unless it is blocked specifically in context. This would also provide a model for overexplicitness in some cross-lingual varieties: Semantic oversaturation occurs if both words in what is expected to be a coselection of one word with another are instead semantically selected, causing the reader to search for

[25] Personally, I am skeptical of any accounts of brain structures in metaphors that claim 'actual' reality, since information appears to be so different from matter after all. On the other hand, usage-based linguistics would not be what it is without a strong cognitive orientation. I prefer to stay agnostic to this question for now, but perhaps it is worthwhile developing graph-based hypotheses and experimental or simulation designs of language-specific cognition in the future.

cues on how to resolve excess semantic weight and cognitive dissonance or confusion.[26]

7.4. Towards a theory of coselectional constraint?

Chapter 2.3 raised some questions with respect to the role of coselection in linguistic theory:

- whether fixed chunks are to be interpreted in the same way as coselectional preferences;
- how coselectional preferences are represented in texts quantiatively and qualitatively (how many, what kinds are there, are they all the same?);
- whether (and how) coselectional preferences are different in learners and native speakers, both quantiatively and qualitatively;
- what can be said about the *structural* role of coselectional preferences or constraints in SLA?

What answers can the insights from this study provide?

Firstly, with a low number of identical coselections overall, the issue of chunks vs. unchunked coselections cannot be meaningfully addressed. While German is a language that would lend itself well to the study of the effects of continuity/discontiguity, it is impossible to tell whether a coselection is memorized as a chunk in a learner or not unless it is repeated frequently or saliently opposed to other structures. For example, if a coselection occurs only in one syntactic environment, while other coselections of the same verb or the same argument are more flexible, this may suggest it is chunked (or primed). However, this would still require frequent reuse, ideally in the same learner, and the Kobalt data do not provide that kind of evidence. Thus, the relationship of word order flexibility and coselection cannot be further illuminated from the present analysis.

It follows also in a more general reply to all of the above questions that coselectional constraints are not easily traced even in homogeneous corpora. Even though section 4.1.2 was able to demonstrate that there is a major lexical overlap between the texts, especially between subcorpora of a language group (over 60% for many BEL texts), there is no evidence for a ubiquity of conventionalized coselection for many of those identical lexemes, certainly not to the extent necessary for an exemplar-based analysis.

To be fair, the data was rather small, especially when further divided by onDaF ranges. Yet, as it was argued earlier, a larger sample may not necessarily provide better evidence due to high combinatorial power: Unless it was larger by at least a magnitude or two, the distribution would likely mostly fill up with more hapax coselections. Also, logically, while more exemplars would be found in larger data, the overall *similarity* between texts should not be affected by larger data.

[26] This is in line with Dux (2016, 426)' observation in reference to Snell-Hornby (1983): "While comparisons with other classes are necessary before arriving at conclusive results, the findings suggest that the descriptivity level (i.e. semantic weight; Snell-Hornby 1983) of a verb class determines the number and nature of its meaning components and valency constructions, as well as the degree to which these differ cross-linguistically. The comparison also revealed unexpected differences in the types of meaning components differentiating individual verbs of diverse semantic classes, and it demonstrated that certain phenomena that are traditionally viewed as independent of verb meaning receive different interpretations when occurring with verbs of different classes."

Identical coselections in Kobalt do exist, and from a qualitative perspective, exemplars are interesting to study: They are functionally and semantically diverse, and many that are lexicalized in the *langue* occur particularly infrequently in the learner texts. But they are clearly not in the vicinity of making up 50-80% of a text as is suggested for chunks in spoken language, and not even in the 20% that was suggested for academic texts by Biber and Conrad (1999).

Chapter 4 has also shown that it is rather difficult to quantify the extent of identical coselections in the first place, because there is no clear object for comparison. Many of the estimations to the extent of fixed language reported in the introduction (chapter 1) are derived from the recognition of formulaic sequences or chunks by native speakers. In other words, native speakers look at a text and decide which parts sound familiar, and then all marked words are divided by the total number of words. Aside from the obvious caveat of confirmation bias, this technique also references a large set of language in which frequency and salience may be conflated – native speakers may recognize something as formulaic because it appears meaningful and familiar, and deduce that it must be frequent.

Or a researcher looks up word combinations in a collocation dictionary, like Nesselhauf (2005) does with 32 essays written by German L1 learners of English. She finds 213 out of 1072 verb + noun combinations in her data to be collocations as defined by a dictionary – this is interesting, but it is not a measure of similarity of texts: If the same number of listed collocations occurs in two texts, that alone is not very informative of their similarity, since not all collocations in the dictionary will possess the same phraseological weight, some may be more frequent, plausible, idiosyncratic, non-compositional, idiomatic…than others; and, depending on the underlying linguistic model, it can make a difference whether learners reuse one of the collocates – for example in several collocations of *haben* ('to have') – or whether they include many rare and diverse verbs and nouns.

A third way of estimating the amount of coselectionally constrained material is to compare continuous chunks (n-grams, for example) across texts, which then contains syntactic artifacts (like *is the*, *the + frequent noun*, etc.). Identical coselections like these do *not* make up one to four fifths of text, at least not in the verb-argument domain in Kobalt. Thus, it appears that on a *parole*-level, there is not an abundance of exemplar-based or item-specific coselectional preferences to measure in Kobalt.

This is puzzling given the high emphasis in all of usage-based linguistics on that phraseology is anything but peripheral, and at the same time the field's insistence on statistical processes as both determining and serving ground for the emergence of linguistic principles and abstractions (Ellis (2008, 2012a); Ellis et al. (2008); Bybee (2013); Bybee and Hopper (2001); Goldberg (2006); Tomasello (2009). In Gries (2014, 45)' summary:

> "Many studies in cognitive/usage-based linguistics have shown that speakers keep track of vast amounts of multidimensional and probabilistic co-occurrence information, and by now it is also well understood how early this begins – in fact, such learning processes begin in utero – and how fast this happens – speakers can pick up meaningless but probabilistically somewhat reliable patterns after just a few minutes of input".

Similarly, Diessel and Hilpert (2016, 17) in review of the study of constructions or collostructions, productivity, syntactic extraction from unknown input, phonetic reduction, segmentation, sentence processing, and markedness conclude: "The research that we have reviewed supports a view of linguistic knowledge in which frequency of use is a funda-

mental determinant of grammatical knowledge" – should identical phrases or identical coselections not make up a larger part of production then?

For L1 one might perhaps argue that the speakers' linguistic in- and output is so inestimably large and varied that coselectional constraint is indeed a statistical epiphenomenon, but one that is never quite traceable in corpora because corpora generally do not reflect a speaker's input (for example, people read one or at best two newspapers, but never as many as are represented in a corpus of newspaper language). It remains for further debate whether this is an valid model of L1 language input or not for most speakers.[27]

However, for a learner, this is hard to argue at all. In fact, a learner reaching a B1 level according to CEFR may have arrived at this point after four semesters of university level teaching, perhaps at four lessons a week – this is at least what is suggested by common practice at Humboldt University of Berlin. With this, they may have had some 240 hours of instruction, about two textbooks worth of vocabulary, and some audio and video material that is unlikely to repeat the same coselections over and over (unless they are very basic), plus whatever time they may spend practicing outside of class, but this is unlikely to happen at large with new material and much more varied input. This material alone appears unlikely to produce statistical results of the fine-grained kind (choosing the lexicalized version out of many possible ones) suggested by the studies cited and this one, unless those are primed or facilitated through another process. This is not to imply a *poverty of the stimulus argument* as it was brought forth by Chomsky and many others for syntax in FLA and SLA (see Cook (1991) for a synopsis of the argument with references). Rather, it marks the observation that with the combinatorial power of words multiplied by their textual contexts, their semantic or valency frames, and their arguments *even as they are found in the input*, finding enough *purely* statistical evidence for forming the right coselections seems unlikely from an early intermediate learner's perspective.

Yet, BEL learners at intermediate stages (BEL-95 might roughly correspond to B1, BEL-115 to early B2) use the most identical coselections out of all Kobalt subcorpora (see section 4), but also have the lowest modularity values, meaning that they at the same time have the lowest number of structured communities in a graph. This suggests that they, at the same time, are more similar to each other, and more random. How could this work?

Rather simply, in fact: It could be simply explained by the distribution of a verb filling up with one prototypical argument (the same for everyone), and then a randomized addition of other arguments (different for everyone), where some of that randomness is later replaced either through a second or a third prototype, or through a whole new verb-argument pair that adds differentiation, and thus structure and modularity to the graph.

A prototype is of course to some degree a frequency observation. But it is not this at its core: A penguin is not less prototypical bird because it is rarer. A bird is also not a more prototypical dinosaur than, say, a brontosaurus or a tyrannosaurus, just because

[27] One argument against this is that if 20 native speakers writing on the same topic do not produce a lot of conventional coselections, and most native speakers cannot be assumed to regularly read even this amount of text on a given topic, from where would the statistical pattern emerge in a way that a native speaker does not perform similarly to a learner for a new topic or register? Or do they? See also Pierrehumbert and Granell (2018, 129) in discussion of morphological productivity: "For the Wikipedia editors, who had a median age of 25 in 2010 (...), reading 8 hours a day from age 5 yields a median estimated exposure to 146,000 alphabetic word types, which is still fewer than the median hapax rank" – meaning that even with this extreme overestimation of input, they could not possibly have encountered more than half of the words they use as a community; and this does not yet account for possible or actual combinations of words.

these days there are more birds than other dinosaurs. Rather, prototypicality emerges from a combined analysis of a category and at least two examplars that are compared to features of the category. As has been argued in chapter 2.2, prototypicality, at its core, is a relational phenomenon, not an ontological one, and categories do not require lots of exemplars to be formed (see for example Lakoff (1987)'s categories of 'things to gift to someone for their birthday', which can be formed ad hoc without having had any experience with that person, and arguably also without having gifted much previously).

Tentatively from the study of coselection here, and with curiosity for further study in this direction, I would suggest the following: There are statistical, or rather, *distributional* aspects of the input, and speakers are sensitive to those. But there are also powerful *linguistic* processes guiding learners and native speakers in coselectional acquisition and in its use. Those processes are both semantic and syntactic, and likely also morphological and phonotactic, where a prototypical coselectional pattern would serve as an anchor for certain concepts in the lexicogrammar of either the target language or its latent structure, the interlanguage in the learner's mind. Some of these anchors may best be understood as notions denoting cultural and linguistic concepts of the target language on different levels of granularity or specificity. A very similar idea has been proposed by Frath and Gledhill (2005, 9):

> "What is of interest is the notion of reference. The test should be: does our expression refer globally to a social object or is it related to other denominators in an on-going discourse? If the latter is true, it is likely that our expression is an instantial, discursive feature of a text, i.e. an interpretant. The collocations strong tea and powerful car refer globally to socially existing complex objects, they are denominators. Powerful tea and strong car do not refer to socially existing objects, and so can only be seen as one-off mistakes or literary creations."[28]

Some coselections may serve as syntactic anchors or even anchors for phonetic regularities (it is well-known that many collocations alliterate). Where syntax is more differentiated, it is also more anchored in coselectional patterns, and it simulataneously allows for more detailed coselectional patterns (rather than randomness). This may be why coselections are harder to memorize than words: Because they map not to a meaning (a concrete extension), but to a concept. A form-meaning pair may be easier to handle than a form-meaning-concept triangle.[29] Where the concept is not yet understood or not fully developed and integrated into the system, it cannot be anchored through a form, explaining why advanced learners find it easier to memorize and remember collocations than those at intermediate stages of target language acquisition.

For example, take the verb *haben* ('to have') that appears frequently in all of Kobalt in the use of *Computer, Handys, Internet haben* ('to have computers, cell phones, the internet'), *Möglichkeiten, Probleme haben* ('to have options/opportunities, problems'), and *Recht, Angst haben* ('to be right, to be afraid'). These three uses are of different kinds, both in terms of their dispersion in the language (*Möglichkeiten, Probleme haben*

[28] This is consistent with a finding from Jolsvai et al. (2013), where participants in a reaction time study shorter reaction times for frequent 3-grams than infrequent ones, but a more determining factor was the meaningfulness of the 3-grams.

[29] This is also congruent with some observations from FLA that suggest that form is easy to pick up, but meaning is hard Naigles (2002).

('to have options/opportunities, problems') is less register- and topic-constrained than *Handys, Computer, Internet haben* ('to have computers, cell phones, the internet')), and in their linguistic typology (compositionality, semantics of the verb, syntactic flexibility – *#Rechte haben, #Ängste haben* ('to have rights, to have fears, insecurities'), but *eine Möglichkeit – zwei Möglichkeiten haben* ('to have one option – two options')). But they still work on the same lexicogrammar. If, as is suggested in emergentist grammars, 'what fires together, wires together' or 'what is used together, fuses together' (Bybee, 2002), *all* of these uses will be connected to the verb *haben*.[30] Yet still they constitute prototypes of different *aspects* of the grammar, for example a possessive vs. a predicational vs. a metaphorically predicational one. These are by the way not all uses that exist at least in Russian (I am unaware of whether they exist in Belarusian): **У неё (есть) страх, *она имеет страх*, ('*u neyo est' strakh – she has fear', '*ona imeet strakh – she owns fear'), rather: *ей страшно* ('ey strashno – to her it is scary'); pointing towards an anchor that is indeed one in the target language system rather than the interlanguage per se.

It is *useful* to know *Recht haben* ('to be right') or *Möglichkeiten haben* ('to have options, opportunities') in German, because they denote culturally defined meaning spaces and linguistic contrasts. There are many ways to express meanings similar to *Möglichkeiten haben* in German, all of them idiomatic: *etwas/verschiedene Dinge tun können* ('to be able to do something/different things'), *Gelegenheit haben* ('to have a chance, opportunity'), *eine große Auswahl haben* ('to have a wide range of choices'), and likely several more. Yet none of them denotes the specific concept or notion of having a possibility + opportunity + option + optimism that *Möglichkeiten haben* does.

How does this go together with the fact that different words with closely related semantics select for different syntactic slots and coselections, as is reported in many studies (see chapter 2.1.3)? There is little conflict, as long as one does not assume that similar semantics of the word sense necessarily anchor in the same structural way. In fact, two words of similar semantics might be helpful for distinguishing between two *syntactic* frames, as in Faulhaber (2011)'s group of *answer, reply* and *respond* group, or in Dux (2016)' frame-semantic analysis of verbs of theft and change in German and English. As such, sometimes, a verb might come with a prototypical frame that might be either more syntactic or more cognitive in kind, like *reply to + indirect object* but *answer + direct object*, while another time, a frame might come with a prototypical verb, like the ditransitive with *give*, but not *donate*. Each of those may be anchoring an aspect of the language, and thereby inherently being more salient, more contrastive to non-prototypical uses, and more lenient to a specific analogy. For example, German *zur Verfügung stehen* ('to be at the disposal (of)') could be viewed as a fancy way of saying *haben* ('to have'), but it also sounds very prototypically German and anchors a support verb construction with a preopositional object, and thus seems richer in several ways.

An anchor can be referred to without explicitly naming it, if reference is obvious by similarity or context. This could explain the observation that idioms and proverbs are in fact rarely found in corpora in their base form (Moon, 1999).[31]

Obviously, this is merely a first sketch of what a functional description of coselection might look like. First steps would require a functional analysis of relevant coselections on

[30] See Schmid (2010) for a similar argument on context-free entrenchment.

[31] Like *kick the bucket*, that is only found in explicit discussions of it in social science lectures, and in fictional prose, in the BNC, with an overall frequency of 7 for the present tense and 6 for the past tense, and some of those being literal uses, not idiomatic.

different linguistic levels and a model of anchoring the right things through the concepts of salience or *noticing* as coined by Schmitt (2004). In constructionist, connectivist, and emergentist models, linguistic concepts are modeled to emerge from 'patterns of usage', implicitly or explicitly relying on frequency and distribution. This has been stated programmatically in opposition to generative approaches, programmatically also taking away all linguistic specificity and relying solely on abstractions over distributions and perhaps cognitive frames. This has brought about a large amount of research showing distributional preferences, but even with the most sophisticated models, whether organized by distributional aspects as in Deshors and Gries (2016)'s mixed-model account of to-infinitive vs. gerund complement, or by semantic aspects, as in Faulhaber (2011) or Dux (2016), is frequency or distribution of usage *alone* unable to explain the phenomenon fully. My suggestion is to accept a spark of linguistic magic back into usage-based linguistics by allowing space in the model for an intrinsically linguistic understanding of speakers of what is helpful, useful, categorially foundational, fundamental to the syntax, or systematic in a language; and for this understanding of course to be coined from use, not as much from frequency as from attention and intention in a system fundamental to cognition and social behavior. This is stated as a positional perspective by the Five Graces Group et al. (2009, abstract):

> "A speaker's behavior is the consequence of competing factors ranging from perceptual constraints to social motivations. The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive mechanisms."

But these aspects have not yet found their way concretely into the modeling of lexicosyntax, and lexicogrammatical studies looking into concrete frequency effects have not yet found their way or merged into a unified theoretical framework of usage-based linguistics either. Perhaps this is because attention, salience, and related concepts are seemingly harder to model than frequencies of co-occurrence. Maybe a reconceptualization of co-occurrences as anchors of different kinds of linguistic understanding can play a constructive role in such a process.

7.5. Summary

In this thesis, findings from usage-based linguistics with respect to the *idiom principle* have been reviewed, concluding that coselectional constraint has been observed as a general tendency of language, but that a theoretical account with high explanatory power has not yet been developed. For a first attempt at the quantification and empirical verification of coselectional constraint as a structural property of natural language, hypotheses based on structural assumptions from usage-based and interlanguage approaches were derived. Core hypotheses include the observability of a process of lexicosyntactic reorganization through the course of SLA, and a lower degree of coselectional constraint in learners vs. native speakers.

A strictly controlled corpus containing texts written by Belarusian (BEL) and Chinese (CH) learners of German and a control group of native speakers (L1) was then processed (digitized, tagged, parsed, manually corrected, and annotated) for a quantitative analysis. A statistical analysis showed that while a predicted process of lexical diversification, randomization and specialization is observable in a number of metrics, an item-based analysis of coselection cannot be performed due to the low frequency and low overlap between items

in subcorpora, and due to the lack of interpretability of any result because there exists no quantification of the *idiom principle* that could serve as a baseline for comparison.

A graph-based model was then suggested as a quantifiable alternative, which is closer to a distributional model of coselection than an item-based. As a quantification, Louvain modularity was applied and yielded results consistent with three main hypotheses: That graphs are more structured in L1 than L2, that graphs are more structured in more advanced learners, and – only for BEL learners – that graphs are least structured at intermediate stages compared to earlier and more advanced stages, consistent with the hypothesized u-shaped development. For CH learners, no robust u-shaped development has been found.

Four aspects were discussed to explain this divergence from the hypotheses: The typological argument that CH learners are likely to prefer verb + noun combinations in vocabulary learning due to the high frequency of verb-noun-compounds in Mandarin Chinese and may exhibit higher coselectional sensitivity due to the frequent necessity of contextual disambiguation in Mandarin. The typological and learning theoretical argument that the two languages might provide a very different framework for the acquisition of verbs and nouns specifically. The cognitive and language environment argument that BEL learners might have overall lower coselectional constraint or be less sensitive to constraining forces because they are proficient bilinguals in a consistently bilingual environment. And the register and metalinguistic argument that learners might respond with different texts to the same prompt for a number of dynamically interacting linguistic and extra-linguistic reasons in discourse, expectation, and skill management.

Large L1 variance has been shown in virtually all analyses, suggesting that in a corpus like this, and perhaps in general, an L1 standard is a vague concept to aspire to in the sense of a specifiable target language space; and that L1 variation requires more attention in corpus research to get a clearer picture of the actual differences between learner and L1 varieties.

In methodological regards, the thesis contributes to the systematization of methodological development in corpus linguistics with a rigorous approach to internal validation through a sliding-window-sampling, an out-of-sampling, and a text length normalization vs. a text-structural analysis. All of these require further theoretical development in the context of corpus linguistics, and replication on unseen data and data from different registers, but so far appear to yield linguistically valuable results and to confirm the results from the grouped analysis.

For external validation, it was stated that replication and extension are necessary. This is true of any study, but of particular relevance since the method was developed on previously seen data, and since a new measure was introduced of which the mechanics are not yet well-understood in application to linguistic research. It has been suggested that an extension to learners of German who are from monolingual Russian areas would be of particular interest, as well as an extension of the Chinese data to match the earlier onset of Belarusian data; But also an extension to other languages with less influence of verb-noun-compounds or a complex verb morphology. It was also suggested that an extension of the method to other registers and L1 contexts is desirable. It was further clarified that only an extension to other learner groups can provide the evidence necessary to decide whether either the BEL or the CH group comply with a norm or whether such a norm exists at all.

With respect to the evaluation of data size as a determining factor, it was suggested that a larger dataset may provide more insight into coselectional exemplars, but would

not necessarily solve the problem of exemplar-based, stratified comparison, since with the long-tailed distribution, a growing number of hapaxes is to be expected. It was also argued that artificially creating a huge dataset of homogeneous data is ontologically and epistemologically confusing in the context of more natural composition of learner language. Instead it was argued that relying on small to medium-sized corpora can provide linguistically grounded insight that is not available from large corpora because those are necessarily linguistically underspecified due to the high cost of annotation of linguistically interesting, and thus typically ambiguous or surface-variable phenomena. With this the development of exact methods that do not require large datasets was argued to be of particular relevance to the study of non-canonical language and subfields that work with inherently sparse data.

In terms of future research, an application of graph-based modeling and graph-theoretical computations to core linguistic issues was suggested. In particular, three problems were defined as worthwhile looking into: Isograph detection, that would help to determine the influence of structural vs. item-specific aspects in lexicosyntax; the definition of closed or interacting subspaces of language in graphs; and modeling of grammar as graph. All of these were presented as possible spaces of quantification where probabilities are problematic concepts, either because they may be a problematic concept in language overall, or because the material that exists to study those fields does not suffice to think in terms of probabilities, for example in less documented languages, language contact situations, or historical linguistics.

Finally, it was argued that coselectional constraints or preferences, being at the heart of lexicogrammar, may play a structural role as prototypical anchors of linguistic concepts in both learners and native speakers. It was suggested that this is not merely a statistical or frequentist effect, but one that combines distributional sensitivity with linguistic insight into structural and/or semantic fundamentals of the language system in question. Some aspects were outlined that might provide a research agenda for a functional theory of coselectional preferences or constraints in the future.

With this, the thesis aspires to be a contribution to the theoretical development of usage-based learner language research, lexicogrammar, and corpus linguistics; as well as a contribution to the systematization of methodological development in quantitative linguistics through a) synthesizing research from usage-based linguistics related to the idiom principle and raising questions towards necessary clarifications of the model; b) showing the limits of statistical measures in the study of coselectional constraint in limited data; c) presenting and validating a new method for the modeling and analysis of lexicogrammar; d) raising questions towards the interpretability of lexicogrammar isolated from questions of higher-order functions such as register, typological, and cultural effects; and e) providing suggestions for future extensions of the graph-based model for quantitative models of aspects of language where abundant data is not to be expected.

The individual chapters show caleidoscopically that coselectional constraint is intertwined with most, and perhaps all, linguistic levels: Lexicon, syntax, semantics, morphology and morpho-phonotactics, pragmatics, text-linguistics; on *langue* and *parole*; and on the interface of corpus linguistics and psycholinguistics. With such richness, nativelike selection is a puzzle not only for linguistic theory (Pawley and Syder, 1983), but perhaps more so for empirical linguistics. This thesis can only humbly provide a first step towards a better understanding of coselectional constraint in learners and native speakers. Its main

contribution thus lies in disentangling some of the complexity and raising questions that may serve as a starting point for future research into the phenomenon.

A. Appendix

A.1. Formal definition of the graph-based model

Formally, the model looks as follows:
The graph G of a subcorpus S is defined as

$$G_S = (V, E, dep, w, doc\_count, sc\_freq, pos, pass, v\_cat) \tag{A.1}$$

where

$$V = \{lexeme_1, ..., lexeme_n\}, \text{ where} \tag{A.2}$$
$$lexeme_a = lexeme_b \text{ iff } lexeme_a = homograph(lexeme_b) \tag{A.3}$$

$$E = \{(lexeme_{source}, lexeme_{target})_1, ..., (lexeme_{source}, lexeme_{target})_n\} \tag{A.4}$$

$$\begin{aligned} dep = E : dep \mapsto \{&ADV, APP, ATTR, AUX, AVZ, CJ, \\ &DET, EXPL, GMOD, GRAD, KOM, KON, KONJ, LOKAL, NEB, \\ &OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, PAR, PART, PN, PP, \\ &PRED, PTKNEG, REL, ROOT, S, SUBJ, SUBJC, VOK, ZEIT\} \end{aligned} \tag{A.5}$$

$$w = E : w \mapsto \mathbb{N} \tag{A.6}$$
$$doc\_count = V : doc\_count \longmapsto \mathbb{N} \tag{A.7}$$
$$sc\_freq = V : sc\_freq \mapsto \mathbb{N} \tag{A.8}$$

$$\begin{aligned} pos = V : pos \mapsto \{&ADJA, ADJD, ADV, APPR, \\ &APPRART, APPO, APZR, ART, CARD, FM, ITJ, KOUI, \\ &KOUS, KON, KOKOMNN, NE, PDS, PDAT, PIS, PIAT, \\ &PIDAT, PPER, PPOSS, PPOSAT, PRELS, PRELAT, RF, \\ &PWS, PWAT, PWAV, PAV, PTKZU, PTKNEG, PTKVZ, \\ &PTKANT, PTKA, TRUNC, VVFIN, VVIMP, VVINF, \\ &VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, \\ &VMFIN, VMINF, VMPP, XY, \$,, \$., \$(\} \end{aligned} \tag{A.9}$$

$$pass = V : pass \mapsto \{T, F, NA\} \tag{A.10}$$

$$v\_cat = V : v\_cat \mapsto \{aux, copula, modal, modifying, particle, prefix, simple, cx, gehen\_cx, mixed\} \tag{A.11}$$

In words: The graph of a subcorpus is a nine-tuple of a set of vertices V, a set of edges E, functions *pos, v_cat* and *pass* that map labels to vertices and function *dep* that maps labels to edges as defined by target sets (Stuttgart-Tübingen Tagset for POS tags on nodes, Schiller et al. (1995) and Foth's dependency grammar tagset for dependency labels on edges, Foth (2006), *TRUE/FALSE* and *NA* for passive, and the verb categorization as defined in the chapter 3.2.2, functions *doc_count* and *sc_freq* that map values from the codomain of natural numbers to vertices, and function w that maps values from the codomain of natural numbers to edges. This is also the *full graph* as defined in the previous section. The property graph notation with properties as functions in the G_S tuple is adopted from Marton et al. (2017) who use it for the definition of neo4j query language Cypher.

Graph specificities *pp, vas_prep, vas_no_prep, no_subj* are subgraphs of G_S, where

$$pp = (G_S, E') \text{ where } E' \subseteq E(G) \text{ such that} \tag{A.12}$$

$$E' = \{(lexeme_{source}, lexeme_{target}) \| \; pos(lexeme_{source}) \in \{VVFIN, VAFIN, VMFIN, VVINF, VAINF, VMINF, VAPP, VMPP, VVPP, VVIZU, APPR, APPRART\}\} \text{ and } dep(E') \subseteq \{AUX, CJ, EXPL, OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, NEB, SUBJ, SUBJC, PP, PN, REL, ROOT, S\} \tag{A.13}$$

$$vas\_prep = (G_S, E') where E' \subseteq E(G) \text{ such that} \tag{A.14}$$

$$E' = \{(lexeme_{source}, lexeme_{target}) \| \; pos(lexeme_{source}) \in \{VVFIN, VAFIN, VMFIN, VVINF, VAINF, VMINF, VAPP, VMPP, VVPP, VVIZU, APPR, APPRART\}\} \text{ and } dep(E') \subseteq \{AUX, CJ, EXPL, OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, NEB, SUBJ, SUBJC, REL, ROOT, S\} \tag{A.15}$$

$$vas\_no\_prep = (G_S, E') where E' \subseteq E(G) \text{ such that} \tag{A.16}$$

$$E' = \{(lexeme_{source}, lexeme_{target}) \| \; pos(lexeme_{source}) \in \{VVFIN, VAFIN, VMFIN, VVINF, VAINF, VMINF, VAPP, VMPP, VVPP, VVIZU\}\}$$
and
$$dep(E') \subseteq \{AUX, CJ, EXPL, OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, NEB, SUBJ, SUBJC, REL, ROOT, S\}$$
and
$$pos(lexeme_{target} | \; lexeme_{target} \in \{(lexeme_{source}, lexeme_{target}) | \; dep((lexeme_{source}, lexeme_{target})) = \{OBJP\}) \in \{NN, NE\} \quad \text{(A.17)}$$

$$no\_subj = (G_S, E') where E' \subseteq E(G) \text{ such that} \quad \text{(A.18)}$$

$$E' = \{(lexeme_{source}, lexeme_{target}) \| \; pos(lexeme_{source}) \in \{VVFIN, VAFIN, VMFIN, VVINF, VAINF, VMINF, VAPP, VMPP, VVPP, VVIZU\}\}$$
and
$$dep(E') \subseteq \{AUX, OBJA, OBJC, OBJD, OBJG, OBJI, OBJP, NEB, SUBJ, SUBJC, REL, S\}$$
and
$$pos(lexeme_{target} | \; lexeme_{target} \in \{(lexeme_{source}, lexeme_{target}) | \; dep((lexeme_{source}, lexeme_{target})) = \{OBJP\}) \in \{NN, NE\}$$
and
$$pass(lexeme_{source} | \; lexeme_{source} \in \{(lexeme_{source}, lexeme_{target}) | \; dep(lexeme_{target}) = SUBJ\}) = T \quad \text{(A.19)}$$

Defining the source nodes as verbs is necessary to exclude the frequent case of an infinitive complement taken by some nouns, such as *Sie haben die Möglichkeit, etwas zu tun* 'they have the chance to do something', where in a graph with unspecified POS in the source node, coselectional constraints over *Möglichkeit* would get lost due to the connection with the much less restricted verb class that can complement it. The embedded VAS *etwas zu tun* 'to do something' is represented in the graph through the verb head *tun*.

A.2. Approximation of the modularity limit

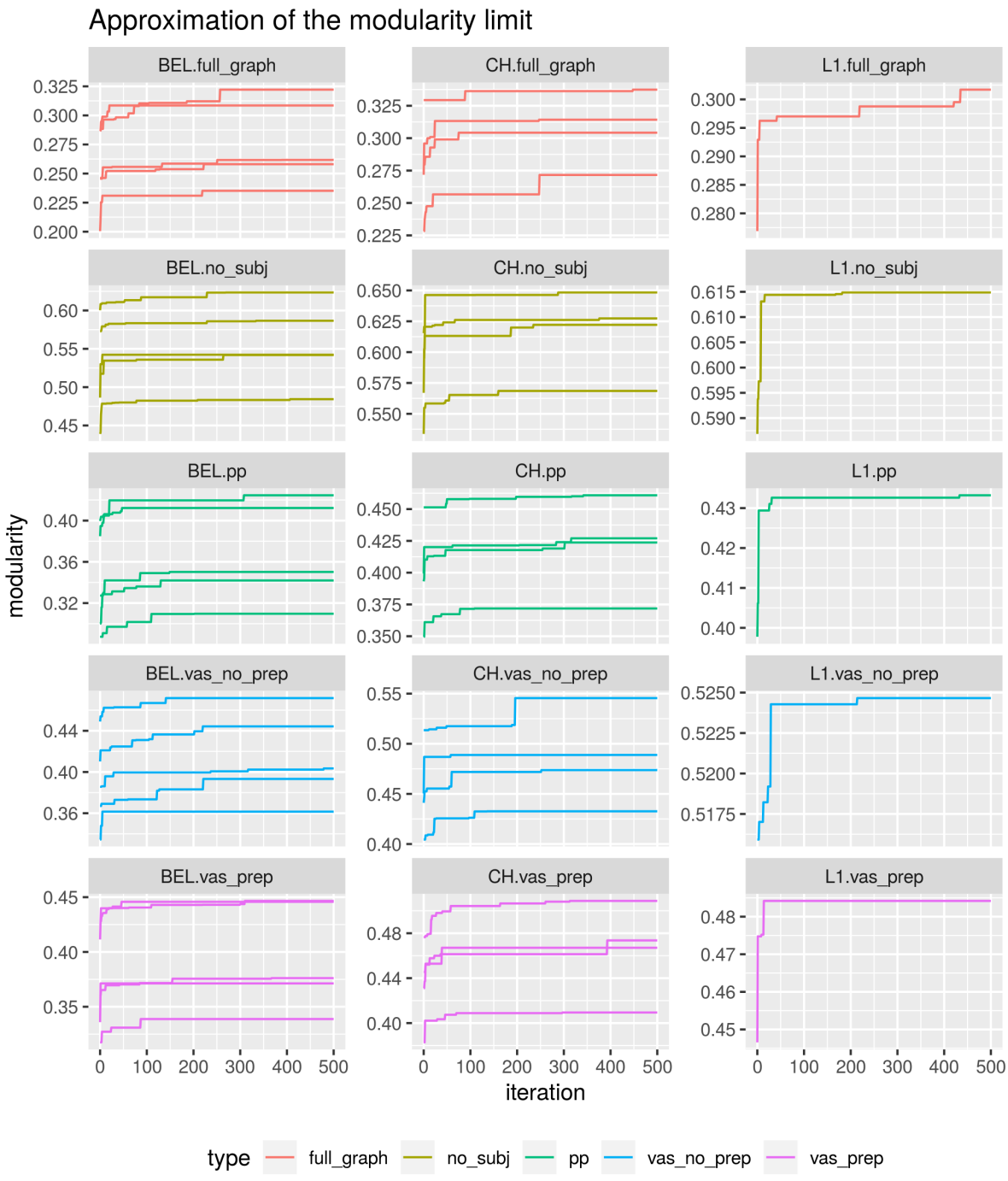


Figure A.1.: Approximation of the modularity limit within 500 iterations, free y-scale. Maximum modularity is reached after around 300 iterations for most graphs. Individual lines represent subcorpora in language groups. More modular graphs appear to reach maximum modularity with higher probability within fewer iterations.

A.3. Weighted vs. unweighted modularity

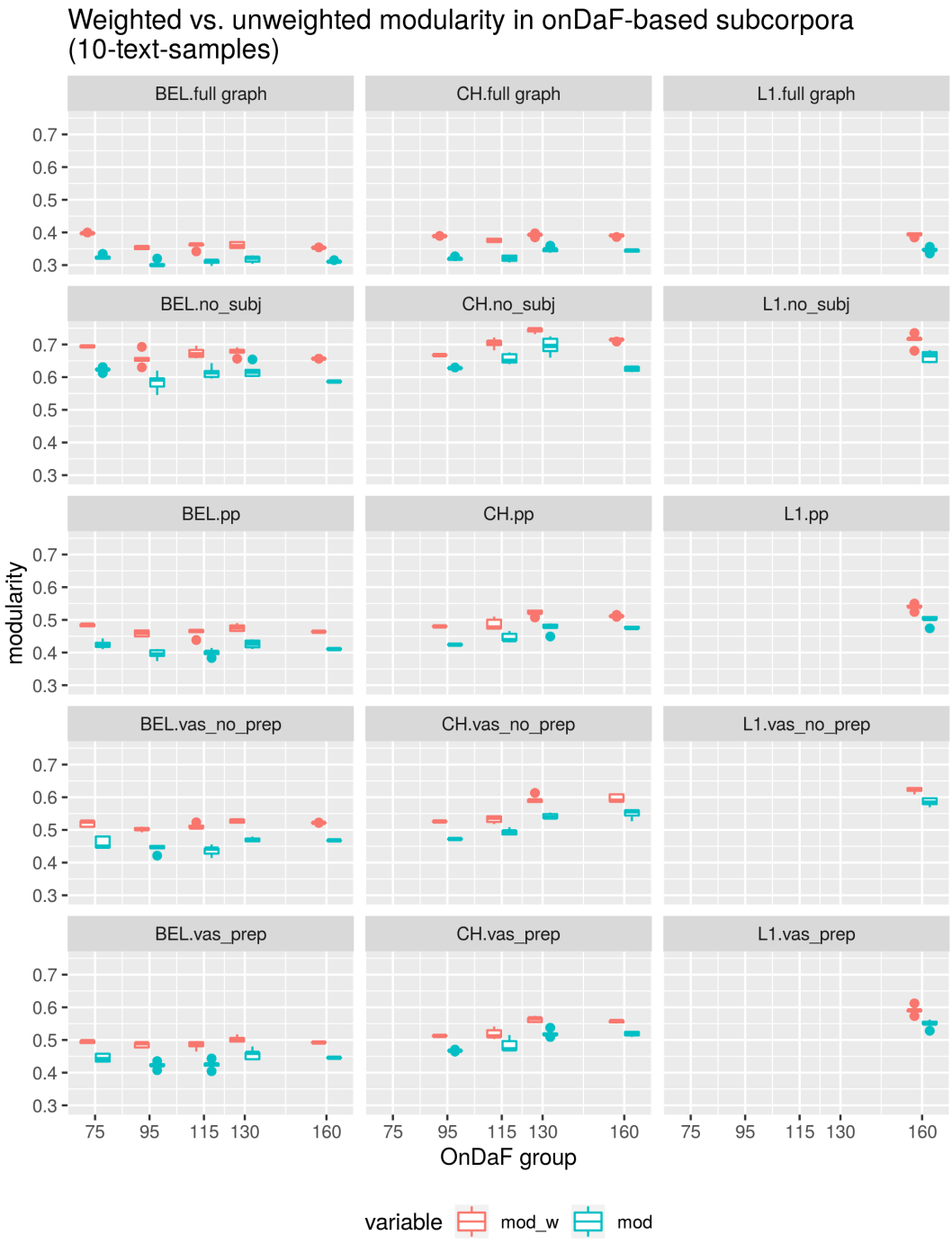


Figure A.2.: Weighted vs. unweighted modularity in onDaF-based grouping

Bibliography

Ágel, V. and Fischer, K. (2010). 50 Jahre Valenztheorie und Dependenzgrammatik. *Zeitschrift für germanistische Linguistik*, 38(2):249–290.

Ahnert, R. and Ahnert, S. E. (2015). Protestant letter networks in the reign of Mary I: a quantitative approach. *ELH*, 82(1):1–33.

Aitchison, L., Corradi, N., and Latham, P. E. (2016). Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS computational biology*, 12(12):e1005110.

Alexandrescu, A. and Kirchhoff, K. (2009). Graph-based learning for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 119–127. Association for Computational Linguistics.

Alishahi, A. and Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive science*, 32(5):789–834.

Alishashi, A. and Stevenson, S. (2005). A probabilistic model of early argument structure acquisition. In *Proceedings of the 27 th Annual Meeting of the Cognitive Science Society*.

Almela, M. (2011). The case for verb-adjective collocations: corpus-based analysis and lexicographical treatment. *Revista de Lingüística y Lenguas Aplicadas*, 6(1):39–52.

Altenberg, B. (1991). Amplifier collocations in spoken English. *English computer corpora: Selected papers and research guide*, 128:133–143.

Ambridge, B., Pine, J. M., Rowland, C. F., and Chang, F. (2012). The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, 88(1):45–81.

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., and Chang, F. (2014). Avoiding dative overgeneralisation errors: semantics, statistics or both? *Language, Cognition and Neuroscience*, 29(2):218–243.

Ambridge, B., Pine, J. M., Rowland, C. F., and Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

Arias-Trejo, N. and Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicon. *Cognition*, 128(2):214–227.

Arnon, I. and Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and speech*, 56(3):349–371.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82.

Aslin, R. N. and Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, 64(s2):86–105.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Baayen, R. H. (2001). *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer Science & Business Media.

Badan, L. (2013). Verb-Object Constructions in Mandarin: a comparison with Ewe. *The linguistic review*, 30(3):373–422.

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT journal*, 47(1):56–63.

Bai, Y., Ding, H., Bian, S., Sun, Y., and Wang, W. (2018). Graph Edit Distance Computation via Graph Neural Networks. *arXiv preprint arXiv:1808.05689*.

Baker, M. C. (1997). Thematic roles and syntactic structure. In Haegeman, L., editor, *Elements of grammar*, pages 73–137. Kluwer, Dordrecht.

Barfield, A. and Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. Springer.

Barrett, M., Kementchedjhieva, Y., Elazar, Y., Elliott, D., and Søgaard, A. (2019). Adversarial Removal of Demographic Attributes Revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Bartsch, S. and Evert, S. (2014). Towards a Firthian notion of collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2:48–61.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154*.

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. (2017). Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*.

Bates, E. and Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In MacWhinney, B., editor, *The emergence of language*, pages 29–79. Lawrence Erlbaum Associates.

Beckage, N., Smith, L., and Hills, T. (2010). Semantic network connectivity is related to vocabulary growth in children. In *Proceedings of CogSci*, volume 10.

Beckert, C. and Juska-Bacher, B. (2015). Bildungssprachliche Kompetenzen bei Schulbeginn: Modellierung-Operationalisierung-Ergebnisse. *Zeitschrift für Literaturwissenschaft und Linguistik*, 45(2):71–89.

Beeching, K. (1997). French for specific purposes: the case for spoken corpora. *Applied linguistics*, 18(3):374–394.

Belz, M. (submitted). *Die Phonetik von äh und ähm*. Dissertation, Humboldt-Universität zu Berlin.

Benson, M. (1989). The structure of the collocational dictionary. *International Journal of Lexicography*, 2(1):1–14.

Bialystok, E., Craik, F., and Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, memory, and cognition*, 34(4):859.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, 26:181–190.

Biber, D. and Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, 15(2):223–250.

Biskup, D. (1992). L1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study. In Arnaud, P. J. L. and Béjoint, H., editors, *Vocabulary and Applied Linguistics*, pages 85–93. Palgrave Macmillan UK, London.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Boas, H. C. (2008a). Determining the structure of lexical entries and grammatical constructions in Construction Grammar. *Annual Review of Cognitive Linguistics*, 6(1):113–144.

Boas, H. C. (2008b). Resolving form-meaning discrepancies in Construction Grammar. In Leino, J., editor, *Constructional reorganization*, pages 11–36. Benjamins, Amsterdam.

Boas, H. C. (2011). A frame-semantic approach to syntactic alternations: The case of build verbs. In Guerrero Medina, P., editor, *Morphosyntactic Alternations in English*, pages 207–234. Equinox, London.

Boas, H. C. (2013). Cognitive Construction Grammar. In Hoffmann, T. and Trousdale, G., editors, *The Oxford Handbook of Construction Grammar*, pages 233–254. Oxford University Press.

Boas, H. C. and Dux, R. (2017). From the past into the present: From case frames to semantic frames. *Linguistics Vanguard*, 3(1):1–14.

Bodomo, A., Yu, S.-s., and Che, D. (2017). Verb-Object Compounds and Idioms in Chinese. In Mitkov, R., editor, *Computational and Corpus-Based Phraseology*, pages 383–396, Cham. Springer International Publishing.

Boers, F., Demecheleer, M., Coxhead, A., and Webb, S. (2014a). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, 18(1):54–74.

Boers, F., Lindstromberg, S., and Eyckmans, J. (2012). Are alliterative word combinations comparatively easy to remember for adult learners? *RELC Journal*, 43(1):127–135.

Boers, F., Lindstromberg, S., and Eyckmans, J. (2014b). Is alliteration mnemonic without awareness-raising? *Language Awareness*, 23(4):291–303.

Borer, H. (1996). Access to Universal Grammar: the real issues. *Behavioral and Brain Sciences*, 19(4):718–720.

Borgatti, S. P. and Halgin, D. S. (2011). On network theory. *Organization science*, 22(5):1168–1181.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. *Language acquisition: The state of the art*, 319:319–346.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., and Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *LREC*, pages 1281–1288.

Boyd, J. K. and Goldberg, A. E. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, 87(1):55–83.

Braine, M. D. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In MacWhinney, B., editor, *Mechanisms of language acquisition*, pages 65–87. Erlbaum, Hillsdale, NJ.

Braine, M. D. S. (1963). On Learning the Grammatical Order of Words. *Psychological Review*, 70(4):323–348.

Brainerd, B. and Chang, S. M. (1982). Number of occurrences in two-state Markov chains, with an application in linguistics. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 225–231.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Braverman, V., Ostrovsky, R., and Zaniolo, C. (2009). Optimal Sampling from Sliding Windows. In *Proceedings of the Twenty-eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '09, pages 147–156, New York, NY, USA. ACM.

Breindl, E. (2011). *Präpositionalobjekte und Präpositionalobjektsätze im Deutschen*. Berlin, Boston: De Gruyter.

Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., and Shein, F. (2015). Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th workshop on multiword expressions*, pages 96–104.

Bueraheng, N. and Laohawiriyanon, C. (2014). Does learners' degree of exposure to English language influence their collocational knowledge? *International Journal of English literature*, 4(3):1–10.

Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, 19(3-4):255–259.

Burger, H. (2004). Phraseologie-Kräuter und Rüben? Traditionen und Perspektiven der Forschung. In Steyer, K., editor, *Wortverbindungen – mehr oder weniger fest*, pages 19–40. Walter de Gruyter GmbH & Co KG.

Butt, M. (2010). The light verb jungle: Still hacking away. In Amberber, M., Baker, B., and Harvey, M., editors, *Complex predicates. Cross-linguistic perspectives on event structure*, pages 48–78. Cambridge University Press Cambridge, MA.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In Givón, T. and Malle, B. F., editors, *The evolution of language out of pre-language*, volume 53 of *Typological Studies in Language*, pages 109–134. Benjamins.

Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions. In Hoffmann, T. and Trousdale, G., editors, *The Oxford Handbook of Construction Grammar*, pages 49–69. Oxford University Press.

Bybee, J. L. and Hopper, P. J., editors (2001). *Frequency and the emergence of linguistic structure*, volume 45 of *Typological Studies in language*. John Benjamins Publishing.

Cai, N., Xue, L., and Fu, Y.-Y. (2015). On Delexicalization Features of Light Verbs in Mandarin. *https://www.semanticscholar.org/paper/On-Delexicalization-Features-of-Light-Verbs-in-Cai-Xue/60f42837c0e71dc4ad006961d629be628ccb5dc4*.

Callahan, D. and Koblenz, B. (1991). Register allocation via hierarchical graph coloring. In *PLDI*, volume 91, pages 192–203.

Callanan, M. A. (1989). Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology*, 25(2):207–216.

Calude, A. S. (2008). Demonstrative clefts and double cleft constructions in spontaneous spoken English*. *Studia Linguistica*, 62(1):78–118.

Camblin, C. C., Gordon, P. C., and Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, 56(1):103–128.

Cantone, K. F. and Haberzettl, S. (2009). „Ich bin dagegen warum sollte man den kein Handy mit nehmen“ - zur Bewertung argumentativer Texte bei Schülern mit Deutsch als Zweitsprache. *Empirische Zugänge zu Spracherwerb und Sprachförderung in Deutsch als Zweitsprache. Münster: Waxmann*, pages 43–65.

Carlucci, L. and Case, J. (2013). On the Necessity of U-Shaped Learning. *Topics in cognitive Science*, 5(1):56–88.

Casenhiser, D. and Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, 8(6):500–508.

Čavar, F. and Tytus, A. E. (2018). Moral judgement and foreign language effect: when the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, 39(1):17–28.

Chan, K. Y. and Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive science*, 34(4):685–697.

Chandy, K. M. and Misra, J. (1982). Distributed computation on graphs: Shortest path algorithms. Technical report, University of Texas at Austin, Department of Computer Science.

Chang, C.-H. (1993). Corpus-based adaptation mechanisms for Chinese homophone disambiguation. *Very large corpora: Academic and industrial perspectives. https://www.aclweb.org/anthology/W93-0300*.

Chen, H., Chen, X., and Liu, H. (2018). How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. *PloS one*, 13(2):e0192545.

Choudhury, M. and Mukherjee, A. (2009). The structure and dynamics of linguistic networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer.

Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735.

Comins, J. A. and Gentner, T. Q. (2015). Pattern-induced covert category learning in songbirds. *Current Biology*, 25(14):1873–1877.

Cong, J. and Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598 – 618.

Conklin, K. and Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied linguistics*, 29(1):72–89.

Cook, P. (2014). Between complex predicates and regular phrases: German collocational clusters. In Müller, S., editor, *Proceedings of the 21st International Conference on Head-Driven Phrase Structure Grammar*, pages 48–62.

Cook, V. J. (1985). Chomsky's universal grammar and second language learning. *Applied linguistics*, 6(1):2–18.

Cook, V. J. (1991). The poverty-of-the-stimulus argument and multicompetence. *Interlanguage Studies Bulletin (Utrecht)*, 7(2):103–117.

Cormode, G., Muthukrishnan, S., Yi, K., and Zhang, Q. (2010). Optimal Sampling from Distributed Streams. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 77–86, New York, NY, USA. ACM.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., and Keysar, B. (2014). Your Morals Depend on Language. *PLOS ONE*, 9(4):1–7.

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Council of Europe (2017). Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors. *Provisional Edition*.

Coupette, C. (2019). *Juristische Netzwerkforschung: Modellierung, Quantifizierung und Visualisierung relationaler Daten im Recht*. Mohr Siebeck.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Crutchley, A. (2007). Comprehension of idiomatic verb+ particle constructions in 6-to 11-year-old children. *First Language*, 27(3):203–226.

Cummings, M. P., Otto, S. P., and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Molecular Biology and Evolution*, 12(5):814–822.

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, 45(3):601–639.

Dąbrowska, E. (2014). Words that go together: Measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon*, 9(3):401–418.

Davis, B. and Maclagan, M. (2010). Pauses, fillers, placeholders and formulaicity in Alzheimer's discourse. In N. Amiridze, B. D. . M. M., editor, *Fillers, pauses and placeholders*, pages 189–215. Benjamins.

De Cock, S., Gilquin, G., and Granger, S. (2009). Introducing LINDSEI, ICLE's talkative sister. In *30th Annual Conference of the International Computer Archive for Modern and Medieval English (ICAME)*.

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., and Navarro, D. J. (2016). Large-scale network representations of semantics in the mental lexicon. *Big data in cognitive science: From methods to insights*, pages 174–202.

De Lacalle, M. L., Laparra, E., and Rigau, G. (2014). Predicate Matrix: extending SemLink through WordNet mappings. In *LREC*, pages 903–909.

Derrible, S. and Kennedy, C. (2011). Applications of graph theory and network science to transit network design. *Transport reviews*, 31(4):495–519.

Deshors, S. C., Götz, S., and Laporte, S. (2016). Linguistic innovations in EFL and ESL. *International Journal of Learner Corpus Research*, 2(2):131–150.

Deshors, S. C. and Gries, S. T. (2016). Profiling verb complementation constructions across New Englishes. *International Journal of Corpus Linguistics*, 21(2):192–218.

Dias, G., Guilloré, S., and Lopes, J. P. (2000). Normalisation of association measures for multiword lexical unit extraction. In *International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*, pages 207–216. Citeseer.

Diessel, H. and Hilpert, M. (2016). Frequency effects in grammar. In *Oxford research encyclopedia of linguistics, E-print: https://www.semanticscholar.org/paper/Frequency-Effects-in-Grammar-Diessel-Hilpert/1b221bb59c4e7cd279621d67965e4ea305ccc8ec*. Oxford University Press.

Diessel, H. and Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*, 12(2):97–141.

Dimroth, C. (2012). Learner varieties. *The encyclopedia of applied linguistics. E-print: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781405198431.wbeal0673*.

Dipert, R. R. (1997). The mathematical structure of the world: The world as graph. *The Journal of Philosophy*, 94(7):329–358.

Divjak, D. and Gries, S. T. (2009). Corpus-based cognitive semantics: A contrastive study of phrasal verbs in English and Russian. *Studies in cognitive corpus linguistics*, pages 273–296.

Dixon, D. (2012). Analysis Tool or Research Methodology: Is there an epistemology for patterns? In *Understanding digital humanities*, pages 191–209. Springer.

Doyle, M. S. (2018). Spanish for the professions and specific purposes: Curricular mainstay. *Hispania*, 100(5):95–101.

During, M. (2016). The dynamics of helping behavior for Jewish refugees during the Second World War: The importance of brokerage. *Online Encyclopedia of Mass Violence*.

Durlauf, S. N. (1993). Nonergodic economic growth. *The Review of Economic Studies*, 60(2):349–366.

Durrant, P. and Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, 47(2):157–177.

Düwel, K. and Heizmann, W. (2009). Einige neuere Publikationen zu den Merseburger Zaubersprüchen: Wolfgang Beck und andere. *Indogermanische Forschungen*, 114:337–356.

Dux, R. J. (2016). *A usage-based approach to verb classes in English and German*. PhD thesis, University of Texas at Austin.

Dębowski, L. (2018). Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy*, 20:85.

Eckes, T. (2010). Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung. *C-test: contributions from current research*, 18:125.

Eckes, T. (2017). Lücken schließen, Brücken bauen: Bestimmung von GER-Niveaus mit dem onSET.

Eckes, T. and Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3):290–325.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 507–509. Association for Computational Linguistics.

Efer, T. (2017). *Graphdatenbanken für die textorientierten e-Humanities*. PhD thesis, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Augustusplatz 10, 04109 Leipzig.

Eimas, P. D. and Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child development*, 65(3):903–917.

El Maarouf, I., Bradbury, J., Baisa, V., and Hanks, P. (2014). Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing. In *LREC*, pages 1001–1006.

Ellis, N. (2012a). Frequency-based accounts of second language acquisition. *The Routledge handbook of second language acquisition*, pages 193–210.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in second language acquisition*, 18(1):91–126.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied linguistics*, 27(1):1–24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, 27(2):164–194.

Ellis, N. C. (2008). The periphery and the heart of language. *Phraseology: An interdisciplinary perspective*, pages 1–13.

Ellis, N. C. (2012b). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of applied linguistics*, 32:17–44.

Ellis, N. C. (2016). Salience, cognition, language complexity, and complex adaptive systems. *Studies in Second Language Acquisition*, 38(2):341–351.

Ellis, N. C. and Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, 93(3):370–385.

Ellis, N. C., O'Donnell, M. B., and Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality.

Ellis, N. C. and Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, 5(1):61–78.

Ellis, N. C., Simpson-Vlach, R., and Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, 42(3):375–396.

Emberson, L. L., Loncar, N., Mazzei, C., Treves, I., and Goldberg, A. E. (2019). The blowfish effect: children and adults use atypical exemplars to infer more narrow categories during word learning. *Journal of child language*, 46(5):938–954.

Emmert-Streib, F., Dehmer, M., and Shi, Y. (2016). Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346-347:180 – 197.

Engel, U. (1996). Tesnière mißverstanden. *Lucien Tesnière – Syntaxe structurale et opérations mentales : Akten des deutsch-französischen Kolloquiums anlässlich der 100. Wiederkehr seines Geburtstages Strasbourg 1993*, pages 53–61.

Engelberg, S. (2014). The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In Boas, H. and Ziem, A., editors, *Constructional Approaches to Syntactic Structures in German*, pages 47–84. De Gruyter Mouton.

Engelberg, S., Meliss, M., Proost, K., and Winkler, E. (2015). *Argumentstruktur zwischen Valenz und Konstruktion*, volume 68. Narr Francke Attempto Verlag.

Erbach, G. and Krenn, B. (1993). *Idioms and support-verb constructions in HPSG*. CLAUS-Report, Computerlinguistik an der Universität des Saarlandes.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1):29–62.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. PhD thesis, Universität Stuttgart.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2):177–190.

Evert, S. (2008). A lexicographic evaluation of German adjective-noun collocations. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*.

Evert, S. and Kermes, H. (2003). Experiments on candidate data for collocation extraction. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Evert, S., Uhrig, P., Bartsch, S., and Proisl, T. (2017). E-VIEW-affilation-A large-scale evaluation study of association measures for collocation identification. *Proceedings of eLex 2017-Electronic lexicography in the 21st century: Lexicography from Scratch*, pages 531–549.

Fang, Z., Schleppegrell, M. J., and Cox, B. E. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research*, 38(3):247–273.

Faulhaber, S. (2011). *Verb valency patterns: A challenge for semantics-based accounts*, volume 71. Walter de Gruyter.

Fehér, O., Ljubičić, I., Suzuki, K., Okanoya, K., and Tchernichovski, O. (2017). Statistical learning in songbirds: from self-tutoring to song culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160053.

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915.

Fillmore, C., Kay, P., and O'Connor, C. (1988). Regularity and idiomaticity in grammatical conditions: The case of LET ALONE. *Language*, 64:501–538.

Fillmore, C. J. (1968). The case for case. In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*, pages 1–88. New York: Holt, Rinehart, and Winston.

Fillmore, C. J. (1977). The case for case reopened. *Syntax and semantics*, 8(1977):59–82.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, volume 1952-59, pages 1–32. The Philological Society, Oxford.

Fischer, A., Suen, C. Y., Frinken, V., Riesen, K., and Bunke, H. (2015). Approximation of graph edit distance based on Hausdorff matching. *Pattern Recognition*, 48(2):331 – 343.

Five Graces Group, Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al. (2009). Language is a complex adaptive system: Position paper. *Language learning*, 59:1–26.

Fornito, A., Zalesky, A., and Bullmore, E. T. (2016). Chapter 9 - Modularity. In Fornito, A., Zalesky, A., and Bullmore, E. T., editors, *Fundamentals of Brain Network Analysis*, pages 303 – 354. Academic Press, San Diego.

Forsberg, F. and Fant, L. (2010). Idiomatically speaking: Effects of task variation on formulaic language in highly proficient users of L2 French and Spanish. *Perspectives on formulaic language: Acquisition and communication, ed. Wood David*, pages 47–70.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75 – 174.

Foster, P., Bolibaugh, C., and Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, 36(1):101–132.

Foth, K. A. (2006). Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. *http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204/*.

Franco, L., Rolls, E. T., Aggelopoulos, N. C., and Jerez, J. M. (2007). Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biological cybernetics*, 96(6):547–560.

Frasca, P., Ravazzi, C., Tempo, R., and Ishii, H. (2013). Gossips and prejudices: Ergodic randomized dynamics in social networks. *IFAC Proceedings Volumes*, 46(27):212–219.

Frath, P. and Gledhill, C. (2005). Free-range clusters or frozen chunks? Reference as a defining criterion for linguistic units. *Recherches Anglaises et Nord Americaines*, 38:25–44.

Fried, M. and Östman, J.-O. (2005). Construction Grammar and spoken language: The case of pragmatic particles. *Journal of pragmatics*, 37(11):1752–1778.

Fuhrhop, N. (2012). *Zwischen Wort und Syntagma: Zur grammatischen Fundierung der Getrennt- und Zusammenschreibung*, volume 513. Walter de Gruyter.

Geeraerts, D. (1989). Introduction: Prospects and problems of prototype theory. *Linguistics*, 27(4):587–612.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2005). Constructions, lexical semantics and the correspondence principle: Accounting for generalizations and subregularities in the realization of arguments. *The syntax of aspect*, pages 215–236.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Goldberg, A. E., Casenhiser, D. M., and Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive linguistics*, 15(3):289–316.

Goldfield, B. A. (2000). Nouns before verbs in comprehension vs. production: the view from pragmatics. *Journal of Child Language*, 27(3):501–520.

Gollan, T. H., Montoya, R. I., Cera, C., and Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of memory and language*, 58(3):787–814.

Golumbic, M. C. (2004). *Algorithmic graph theory and perfect graphs*, volume 57. Elsevier.

Gonzalez, J. E., Low, Y., Gu, H., Bickson, D., and Guestrin, C. (2012). Powergraph: Distributed graph-parallel computation on natural graphs. In *Presented as part of the 10th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 12)*, pages 17–30.

Goodman, N. (1961). *Graphs for linguistics*. American Mathematical Society.

Granger, S. (2005). Pushing back the limits of phraseology: How far can we go. In *Proceedings of the Phraseology 2005 Conference*, pages 1–4.

Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. *Oslo Studies in Language*, 9(3).

Granger, S. and Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52(3):229–252.

Granger, S. and Bestgen, Y. (2017). Using collgrams to assess L2 phraseological development: A replication study. *Language, Learners and Levels: Progression and Variation. Louvain-la-Neuve: Presses universitaires de Louvain*, pages 385–408.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). International corpus of learner English.

Granger, S. and Meunier, F. (2008). *Phraseology: An interdisciplinary perspective*. John Benjamins Publishing.

Greenbaum, E. S. (2014). The development of the International Corpus of English. In *English corpus linguistics*, pages 95–104. Routledge.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Chicago Linguistic Society*, volume 35, pages 151–166.

Gries, S. (2014). Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us. *Developments in English: Expanding electronic evidence*, pages 29–47.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1):137–166.

Gries, S. T. (2019). 15 years of collostructions. *International Journal of Corpus Linguistics*, 24(3):385–412.

Gries, S. T. and Adelman, A. S. (2014). Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In *Yearbook of Corpus Linguistics and Pragmatics 2014*, pages 35–54. Springer.

Gries, S. T. and Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, 9(1):109–136.

Gries, S. T. and Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective onalternations'. *International journal of corpus linguistics*, 9(1):97–129.

Gries, S. T. and Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, 3(1):182–200.

Gries, S. T. and Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, 7(1):163–186.

Guerrero, M. D. (2004). Acquiring academic English in one year: An unlikely proposition for English language learners. *Urban Education*, 39(2):172–199.

Güngör, F. and Uysal, H. H. (2016). A Comparative Analysis of Lexical Bundles Used by Native and Non-native Scholars. *English Language Teaching*, 9(6):176–188.

Gutzmann, M. (2017). Bildungssprache - auch im Fachunterricht. *Grundschule aktuell*, 137:6–8.

Guz, E. (2017). Refining the methodology for investigating the relationship between fluency and the use of formulaic language in learner speech. *Research in Language*, 14(2):[95]–122.

Haas, T. C. (1990). Lognormal and Moving Window Methods of Estimating Acid Deposition. *Journal of the American Statistical Association*, 85(412):950–963.

Haberzettl, S. (2009). Förderziel: Komplexe Grammatik. *Zeitschrift für Literaturwissenschaft und Linguistik*, 39(1):80–95.

Haberzettl, S. (2016). Bildungssprache im Kontext von Mehrsprachigkeit. Eine Untersuchung von Berichtstexten ein-und mehrsprachiger Schüler. *Diskurs Kindheits-und Jugendforschung/Discourse. Journal of Childhood and Adolescence Research*, 11(1):61–80.

Haentjens Dekker, R. and Birnbaum, D. (2017). It's more than just overlap: Text As Graph. In *Proceedings of Balisage: The Markup Conference 2017*. Balisage Series on Markup Technologies, vol. 19.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Halliday, M. A. (1992). Language as system and language as instance: The corpus as a theoretical construct. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, pages 61–77.

Hampe, B. (2011). Discovering constructions by means of collostruction analysis: The English denominative construction. *Cognitive Linguistics*, 22-2:211–245.

Han, Z. and Tarone, E. (2014). Interlanguage: Forty years later (Vol. 39).

Hancioğlu, N., Neufeld, S., and Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, 27(4):459–479.

Handwerker, B. and Madlener, K. (2009). *Chunks für DaF: theoretischer Hintergrund und Prototyp einer multimedialen Lernumgebung*. Schneider-Verlag Hohengehren.

Harary, F. (1959). Graph theory and electric networks. *IRE Transactions on Information Theory*, 5(5):95–109.

Harris, R. A. (1995). *The linguistics wars*. Oxford University Press on Demand.

Hashimoto, B. J. and Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, 69(4):839–872.

Haslhofer, B., Isaac, A., and Simon, R. (2018). Knowledge Graphs in the Libraries and Digital Humanities Domain. *CoRR*, abs/1803.03198.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4(2):237–258.

Hee, K. (2017). Differenzierter Sprachgebrauch in schulischen Interaktionsformen. *Bulletin VALS-ASLA*, 2:115–131.

Hee, K. (2019). Usuelle Wortverbindungen in schulischen Kontexten. *Linguistik online*, 96(3):63–92.

Hentschel, G. (2014). Belarusian and Russian in the mixed speech of Belarus. *Congruence in contact-induced language change: Language families, typological resemblance, and perceived similarity*, pages 93–121.

Herbst, T. (2014a). Idiosyncrasies and generalizations: Argument structure semantic roles and the valency realization principle. *Yearbook of the German Cognitive Linguistics Association*, 2(1):253–290.

Herbst, T. (2014b). The valency approach to argument structure constructions. *Constructions-collocations-patterns*, pages 167–216.

Herbst, T. and Uhrig, P. (2009). Erlangen Valency Pattern Bank-a corpus-based research tool for work on valency and argument structure constructions. Website.

Hilles, S. (1991). Access to Universal Grammar in second language acquisition. *Point counterpoint: Universal Grammar in the second language*, pages 305–338.

Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, 2(2):243–256.

Hilpert, M. (2012). Diachronic collostructional analysis meets the noun phrase. *The Oxford handbook of the history of English*.

Hilpert, M. (2017). Frequencies in diachronic corpora and knowledge of language. *The changing English language—Psycholinguistic perspectives*, pages 49–68.

Hirschmann, H. (2015). *Modifikatoren im Deutschen: ihre Klassifizierung und varietätenspezifische Verwendung*. Stauffenburg-Verlag.

Hirschmann, H., Lüdeling, A., Rehbein, I., Reznicek, M., and Zeldes, A. (2013). Underuse of syntactic categories in Falko. A case study on modification. *Granger, S., Gilquin, G. und Meunier, F.(Hgg.), Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*, pages 223–234.

Hoey, M. (2004). Textual colligation: a special kind of lexical priming. In *Advances in corpus linguistics*, pages 169–194. Brill Rodopi.

Hoey, M. (2012). *Lexical priming: A new theory of words and language*. Routledge.

Holland, J. H., Gong, T., Minett, J., Ke, J., and Wang, W. (2005). Language acquisition as a complex adaptive system. *Language acquisition, change and emergence*, pages 411–435.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied linguistics*, 19(1):24–44.

Hudson, R. and Hudson, R. A. (2007). *Language networks: The new word grammar*. Oxford University Press.

Hughlings Jackson, J. (1874). On the nature of the duality of the brain. In Taylor, J., editor, *Selected writings of John Hughlings jackson, 1958*, volume II, pages 129–145. London: Staples Press.

Hunston, S. (2012). *Pattern Grammar*. Blackwell Publishing Ltd.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Ighreiz, A., Rolfes, L., Shadrova, A., Tischbirek, A., and Möllers, C. (in prep.). *Karlsruher Kanones? Zur Selbst- und Fremdkanonisierung des BVerfG*.

Imai, M., Haryu, E., and Okada, H. (2005). Mapping novel nouns and verbs onto dynamic action events: Are verb meanings easier to learn than noun meanings for Japanese children? *Child development*, 76(2):340–355.

Imo, W. (2011). Die Grenzen von Konstruktionen: Versuch einer granularen Neubestimmung des Konstruktionsbegriffs der Construction Grammar. *Sprachliches Wissen zwischen Lexikon und Grammatik*, pages 113–147.

Ivanova, I. and Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta psychologica*, 127(2):277–288.

Jain, A. and Chang, E. Y. (2004). Adaptive Sampling for Sensor Networks. In *Proceeedings of the 1st International Workshop on Data Management for Sensor Networks: In Conjunction with VLDB 2004*, DMSN '04, pages 10–16, New York, NY, USA. ACM.

Jarvella, R. J. and Sinnott, J. (1972). Contextual constraints on noun distributions to some English verbs by children and adults. *Journal of Verbal Learning and Verbal Behavior*, 11(1):47–53.

Jaworska, S. (2015). Review of recent research (1998-2012) in German for Academic Purposes (GAP) in comparison with English for Academic Purposes (EAP): cross-influences, synergies and implications for further research. *Language Teaching*, 48(2):163–197.

Jelinek, F., Bahl, L., and Mercer, R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256.

Jiang, N. A. and Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3):433–445.

Johnson, C. R., Schwarzer-Petruck, M., Baker, C. F., Ellsworth, M., Ruppenhofer, J., and Fillmore, C. J. (2003). Framenet: Theory and practice. *http://nbn-resolving.de/urn:nbn:de:bsz:mh39-54163*, page 80.

Jolsvai, H., McCauley, S. M., and Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, pages 692–697. Austin, TX: Cognitive Science Society.

Kapustin, V. and Jamsen, A. (2007). Vertex degree distribution for the graph of word co-occurrences in Russian. In *Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing*, pages 89–92.

Kärchner-Ober, R., Hunger, A., and Werner, S. (2015). German for specific purposes (GSP)-a pathway to studies in engineering at University of Duisburg-Essen. In *Proceedings of the 43rd SEFI Annual Conference*.

Kay, P. (2005). Argument structure constructions and the argument-adjunct distinction. *Grammatical constructions: Back to the roots*, 4:71–98.

Ke, J. (2007). Complex networks and human language. *arXiv preprint cs/0701135*.

Kecskes, I. (2007). Formulaic language in English lingua franca. *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects*, 1:191–218.

Kecskes, I. (2015). Is the idiom principle blocked in bilingual L2 production. *Bilingual figurative language processing*, 28:53.

Kempe, V., Gauvrit, N., and Forsyth, D. (2015). Structure emerges faster during cultural transmission in children than in adults. *Cognition*, 136:247–254.

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., and Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts*, 10(4):377.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *Tesol Quarterly*, 37(3):467–487.

Kerbel, D. and Grunwell, P. (1997). Idioms in the classroom: an investigation of language unit and mainstream teachers' use of idioms. *Child Language Teaching and Therapy*, 13(2):113–123.

Khateb, A., Shamshoum, R., and Prior, A. (2017). Modulation of language switching by cue timing: Implications for models of bilingual language control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(8):1239.

Kidd, C., White, K. S., and Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, 14(4):925–934.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

Kilgarriff, A. and Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. In *Proc. Collocations workshop. ACL 2001, Toulouse*, pages 32–38. Citeseer.

Kittel, B., Lindner, D., Tesch, S., and Hentschel, G. (2010). Mixed language usage in Belarus: the sociostructural background of language choice.

Klein, W. (1991). Seven trivia of language acquisition. *Point counterpoint: Universal grammar in the second language*, pages 49–69.

Klein, W. (1998). The Contribution of Second Language Acquisition Research. *Language Learning*, 48(4):527–549.

Klein, W. and Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). *Second language research*, 13(4):301–347.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Kleinschmidt, D. F. and Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*, 23(3):678–691.

Klibanoff, R. S. and Waxman, S. R. (2000). Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children. *Child development*, 71(3):649–659.

Kobyliński, Ł. and Przepiórkowski, A. (2008). Definition extraction with balanced random forests. In *International Conference on Natural Language Processing*, pages 237–247. Springer.

Koch, P. and Oesterreicher, W. (1985). *Sprache der Nähe—Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte*. Walter de Gruyter.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koolen, M. and Kamps, J. (2009). What's in a link? from document importance to topical relevance. In *Conference on the Theory of Information Retrieval*, pages 313–321. Springer.

Koplenig, A. (2017). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*.

Krahmer, E., Erk, S. v., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Krause, T. (2019). *ANNIS: A graph-based query system for deeply annotated text corpora*. PhD Thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

Krenn, B. (2000). Empirical implications on lexical association measures. In *Proceedings of The Ninth EURALEX International Congress*.

Krenn, B., Evert, S., et al. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46.

Krumke, S. O. and Noltemeier, H. (2005). Netzwerkdesign und Routing. In *Graphentheoretische Konzepte und Algorithmen*, pages 299–316. Springer.

Kuczera, A. (2017). Graphentechnologien in den Digitalen Geisteswissenschaften. *ABI Technik*, 37(3):179–196.

Kuiper, K., Columbus, G., and Schmitt, N. (2009). The acquisition of phrasal vocabulary. In *Language acquisition*, pages 216–240. Springer.

Kuno, S. and Takami, K.-i. (2004). *Functional Constraints in Grammar: On the unergative–unaccusative distinction*. John Benjamins.

Kupietz, M., Lüngen, H., Kamocki, P., and Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. *Chicago: University of Chicago*.

Lakoff, G. (1999). Cognitive models and prototype theory. *Concepts: Core Readings*, pages 391–421.

Lambert, C. and Kormos, J. (2014). Complexity, Accuracy, and Fluency in Task-based L2 Research: Toward More Developmentally Based Measures of Second Language Acquisition. *Applied Linguistics*, 35(5):1–9.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Larsen-Freeman, D. (2006). Second language acquisition and the issue of fossilization: There is no end, and there is no state. *Studies of fossilization in second language acquisition*, pages 189–200.

Lasch, A. and Ziem, A. (2014). *Grammatik als Netzwerk von Konstruktionen: Sprachwissen im Fokus der Konstruktionsgrammatik*, volume 15. Walter de Gruyter.

Lau, J. H., Baldwin, T., and Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):10.

Laufer, B. and Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2):647–672.

Lazos, L., Poovendran, R., Meadows, C., Syverson, P., and Chang, L. (2005). Preventing wormhole attacks on wireless ad hoc networks: a graph theoretic approach. In *IEEE Wireless Communications and Networking Conference, 2005*, volume 2, pages 1193–1199. IEEE.

Leibniz-Institut für Deutsche Sprache (2019). Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2019-I (Release vom 18.03.2019).

Lenz, B. (1993). Probleme der Kategorisierung deutscher Partizipien. *Zeitschrift für Sprachwissenschaft*, 12(1):39–76.

Lerner, R. M. (2012). Developmental science: Past, present, and future. *International Journal of Developmental Science*, 6(1-2):29–36.

Leshchenko, J., Dotsenko, T., and Ostapenko, T. (2018). Cross-Linguistic Collocations Used by Bilingual Native Speakers-A Case Study of Komi-Permyak-Russian Bilinguals. *Athens Journal of Philology*, 5(4):301–316.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levin, B., Hovav, M. R., and Keyser, S. J. (1995). *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press.

Leyton, C. E., Savage, S., Irish, M., Schubert, S., Piguet, O., Ballard, K. J., and Hodges, J. R. (2014). Verbal repetition in primary progressive aphasia and Alzheimer's disease. *Journal of Alzheimer's Disease*, 41(2):575–585.

Li, Y., Wei, L., Niu, Y., and Yin, J. (2005). Structural organization and scale-free properties in Chinese Phrase Networks. *Chinese Science Bulletin*, 50(13):1305–1309.

Lieberman, E., Hauert, C., and Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, 433(7023):312.

Lieven, E. V., Pine, J. M., and Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of child language*, 24(1):187–219.

Lindholm, C. and Wray, A. (2011). Proverbs and formulaic sequences in the language of elderly people with dementia. *Dementia*, 10(4):603–623.

Lindstromberg, S., Eyckmans, J., and Connabeer, R. (2016). A modified dictogloss for helping learners remember L2 academic English formulaic sequences for use in later writing. *English for specific purposes*, 41:12–21.

Liu, Z., Wang, H., Wu, H., and Li, S. (2010). Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 825–833. Association for Computational Linguistics.

López, L. S. (2015). An analysis of the integration of service learning in undergraduate Spanish for specific purposes programs in higher education in the United States. *Cuadernos de ALDEEU*, 28(1):155–170.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. *Fortgeschrittene Lernervarietäten*, pages 119–140.

Lüdeling, A., Hirschmann, H., and Shadrova, A. (2017). Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning*, 67(S1):96–129.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, 1:14–17.

Lupu, Y. and Voeten, E. (2012). Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights. *British Journal of Political Science*, 42(2):413–439.

L'homme, M.-C. and Bertrand, C. (2000). Specialized lexical combinations: should they be described as collocations or in terms of selectional restrictions. In *Proceedings. Ninth EURALEX International Congress*, pages 497–506.

Maclagan, M., Davis, B., and Lunsford, R. (2008). Fixed expressions, extenders and metonymy in the speech of people with Alzheimer's disease. *Phraseology: An interdisciplinary perspective*, Not in series(139):175–187.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of child language*, 31(4):883–914.

MacWhinney, B. (2014). Item-based patterns in early syntactic development. *Constructions, collocations, patterns*, 2562:33–69.

Malmkjaer, K. (1993). Who can make nice a better word than pretty. *Text and Technology. In Honor of John Sinclair*, pages 213–232.

Mandke, K., Meier, J., Brookes, M. J., O'dea, R. D., Van Mieghem, P., Stam, C. J., Hillebrand, A., and Tewarie, P. (2018). Comparing multilayer brain networks between groups: Introducing graph metrics and recommendations. *NeuroImage*, 166:371–384.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Markman, A. B. and Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):54.

Martin, E. (2010). Designing and Implementing a French-for-Specific-Purposes (FSP) program: lessons learned from ESP. *Global Business Languages*, 5(1):3.

Martin, R. and Sunley, P. (2012). The place of path dependence in an evolutionary perspective on the economic landscape. In Boschma, R. and Martin, R., editors, *The Handbook Of Evolutionary Economic Geography*. Edward Elgar.

Martin, W. (2008). A unified approach to semantic frames and collocational patterns. *Phraseology: an interdisciplinary perspective*, page 51.

Martinčić-Ipšić, S., Margan, D., and Meštrović, A. (2016). Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Physica A: Statistical Mechanics and its Applications*, 457:117 – 128.

Marton, J., Szárnyas, G., and Varró, D. (2017). Formalising openCypher graph queries in relational algebra. In *European Conference on Advances in Databases and Information Systems*, pages 182–196. Springer.

Massip-Bonet, À. (2013). Language as a Complex Adaptive System: Towards an Integrative Linguistics. In Massip-Bonet, À. and Bastardas-Boada, A., editors, *Complexity Perspectives on Language, Communication and Society*, pages 35–60. Springer Berlin Heidelberg, Berlin, Heidelberg.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory*, 5(1):79–103.

McKay, B. D. and Piperno, A. (2014). Practical graph isomorphism, II. *Journal of Symbolic Computation*, 60:94–112.

McRae, K., Ferretti, and Liane Amyote, T. R. (1997). Thematic roles as verb-specific concepts. *Language and cognitive processes*, 12(2-3):137–176.

Medaglia, J. D., Ramanathan, D. M., Venkatesan, U. M., and Hillary, F. G. (2011). The challenge of non-ergodicity in network neuroscience. *Network: Computation in Neural Systems*, 22(1-4):148–153.

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. *Corpus linguistics. An international handbook of the science of language and society*, pages 328–382.

Mehler, A., Lücking, A., Banisch, S., Blanchard, P., and Job, B. (2016). *Towards a theoretical framework for analyzing complex linguistic networks*. Springer.

Mel'cuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102.

Mendoza, A. and Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, 35:41–55.

Mervis, C. B. and Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, pages 258–266.

Mesarovic, M. D. (1964). Foundations for a general systems theory. In *Proceedings of the Second Systems Symposium at Case Institute of Technology: Views on general systems theory*, pages 1–24. Nueva York, John Wiley & Sons.

Mesbahi, M. (2002). On a dynamic extension of the theory of graphs. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, volume 2, pages 1234–1239. IEEE.

Meunier, F. and Granger, S. (2008). *Phraseology in foreign language learning and teaching*. John Benjamins Publishing.

Michaelis, L. A. (2012). Making the case for construction grammar. *Sign-based construction grammar*, pages 31–68.

Mihalcea, R. and Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge university press.

Mislevy, R. J. and Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning*, 59:249–267.

Molenaar, P. C. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 50(1):60–69.

Mollet, E., Wray, A., and Fitzpatrick, T. (2012). Accessing second-order collocation through lexical co-occurrence networks. *The Phraseological View of Language: A Tribute to John Sinclair*, page 87.

Moon, R. (1999). Needles and haystacks, idioms and corpora: Gaining insights into idioms, using corpus analysis. *The perfect learners' dictionary*, pages 265–281.

Müller, A., Geist, B., and Grimm, A. (2016). (Vor-) Schulkinder mit Deutsch als Zweitsprache im Fokus von Spracherwerbsforschung und Sprachdidaktik. *Diskurs Kindheits-und Jugendforschung/Discourse. Journal of Childhood and Adolescence Research*, 11(1):3–7.

Müller, S. (2002). *Complex predicates: Verbal complexes, resultative constructions, and particle verbs in German*, volume 13. CSLI publications Stanford.

Müller, S. (2010). *Grammatiktheorie*. Stauffenburg.

Müller, S. (2013a). *Head-Driven Phrase Structure Grammar: Eine Einführung*. Stauffenberg.

Müller, S. (2013b). Unifying everything: Some remarks on simpler syntax, construction grammar, minimalism, and HPSG. *Language*, pages 920–950.

Müller, S. (2017). Head-Driven Phrase Structure Grammar, Sign-Based Construction Grammar, and Fluid Construction Grammar. *Constructions and Frames*, 9(1):139–173.

Müller, S. and Wechsler, S. (2014). Lexical approaches to argument structure. *Theoretical Linguistics*, 40(1-2):1–76.

Myles, F. (2004). From data to theory: The over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society*, 102(2):139–168.

Myles, F., Hooper, J., and Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreignlanguage learning. *Language learning*, 48(3):323–364.

Myles, F., Mitchell, R., and Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in second language acquisition*, 21(1):49–80.

Nagy, W. and Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1):91–108.

Naigles, L. R. (2002). Form is easy, meaning is hard: Resolving a paradox in early child language. *Cognition*, 86(2):157–199.

Namy, L. L., Campbell, A. L., and Tomasello, M. (2004). The changing role of iconicity in nonverbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development*, 5(1):37–57.

Nastase, V. (2008). Unsupervised all-words word sense disambiguation with grammatical dependencies. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, 24(2):223–242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Amsterdam.

Newman, I., Benz, C. R., and Ridenour, C. S. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum*. SIU Press.

Nguyen, T. M. H. and Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, 21(3):298–320.

Nicolle, S. (2009). Go-and-V, come-and-V, go-V and come-V: A corpus-based account of deictic movement verb constructions. *English Text Construction*, 2(2):185–208.

Niglas, K. (2007). Introducing the quantitative-qualitative continuum: an alternative view on teaching research methods courses. *Learning and teaching of research methods at university*, pages 185–203.

Nippold, M. A., Moran, C., and Schwarz, I. E. (2001). Idiom understanding in preadolescents: Synergy in action. *American Journal of Speech-Language Pathology*.

Nivre, J. (2004). Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57. Association for Computational Linguistics.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16. Springer.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Noe, C. (2003). French for Specific Purposes: One-size or tailor-made courses? *Mediating between theory and practice in the context of different learning cultures and languages.-[d.: Newby D.].-Graz: Council of Europe.-2003.-* , pages 195–202.

Ogden, C. K. and Richards, I. A. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, volume 29. K. Paul, Trench, Trubner & Company, Limited.

O'Grady, W. (1996). Language acquisition without Universal Grammar: a general nativist proposal for L2 learning. *Second Language Research*, 12(4):374–397.

Ohmori, K. and Higashida, M. (1999). Extracting bilingual collocations from non-aligned parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI99)*, pages 88–97. Citeseer.

Onwuegbuzie, A. J. and Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International journal of social research methodology*, 8(5):375–387.

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]*.

Orliac, B. and Dillinger, M. (2003). Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, pages 292–298.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, 24(4):492–518.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, 63:1–24.

Owoeye, S. T. (2010). Optimal Activation of French for Specific Purposes for Human Development in Nigeria. *Applied Social Dimensions of Language Use and Teaching in West Africa*, pages 224–230.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Palomino-Garibay, A., Camacho-Gonzalez, A. T., Fierro-Villaneda, R. A., Hernandez-Farias, I., Buscaldi, D., Meza-Ruiz, I. V., et al. (2015). A random forest approach for authorship profiling. In *Proceedings of CLEF*.

Pålsson Syll, L. (2012). Rational expectations: A fallacious foundation for macroeconomics in a non-ergodic world. *Real-world economics review; 62*.

Pan, F., Reppen, R., and Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, 21:60–71.

Papo, D. (2013). Why should cognitive neuroscientists study the brain's resting state? *Frontiers in human neuroscience*, 7:45.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics*, 18(3):391–417.

Paquot, M. (2015). Lexicography and phraseology. *The Cambridge handbook of corpus linguistics*, pages 460–477.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, 15(1):29–43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1):121–145.

Paquot, M. and Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32:130–149.

Paquot, M., Naets, H., and Gries, S. (to appear). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb+ object structures in LONGDALE. In Le Bruyn, B. and Paquot, M., editors, *Second Language Acquisition and Learner Corpora*. Cambridge University Press.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pawley, A. et al. (2007). Developments in the study of formulaic language since 1970: A personal view. *Phraseology and culture in English*, pages 3–45.

Pawley, A. and Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication*, 191:225.

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the workshop on WordNet and other lexical resources, second meeting of the north american chapter of the association for computational linguistics*, pages 41–46.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

Perek, F. and Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, 168:276–293.

Perkins, K., Brutten, S. R., and Gass, S. M. (1996). An investigation of patterns of discontinuous learning: implications for ESL measurement. *Language Testing*, 13(1):63–82.

Petersen, I. (2014). „Das von ihnen dargestellte Problem zur Leistungsbewertung in den Schulen "-komplexe Nominalphrasen in Texten von Schüler/innen und Studierenden mit Deutsch als Erst-und Zweitsprache. *Zweitspracherwerb im Jugendalter*, 4:125.

Petrović, S., Šnajder, J., and Bašić, B. D. (2010). Extending lexical association measures for collocation extraction. *Computer Speech & Language*, 24(2):383–394.

Pfaff, C., Traugott, E., LaBrum, R., and Shepherd, S. (1980). Acquisition and development of 'Gastarbeiterdeutsch'by migrant workers and their children in Germany. In *Papers from the Fourth International Conference on Historical Linguistics (Amsterdam)*, volume 38, pages 1–95.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Pierrehumbert, J. and Granell, R. (2018). On Hapax Legomena and Morphological Productivity. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 125–130, Brussels, Belgium. Association for Computational Linguistics.

Pitzl, M.-L. (2012). Creativity meets convention: Idiom variation and remetaphorization in ELF. *Journal of English as a Lingua Franca*, 1(1):27–55.

Plank, F. (1984). Verbs and objects in semantic agreement: Minor differences between English and German that might suggest a major one. *Journal of Semantics*, 3(4):305–360.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *The Modern Language Journal*, 98(1):450–470.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

Pons, F. (2006). The effects of distributional learning on rats' sensitivity to phonetic information. *Journal of Experimental Psychology: Animal Behavior Processes*, 32(1):97.

Prior, A. and MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and cognition*, 13(2):253–262.

Prior, A., MacWhinney, B., and Kroll, J. F. (2007). Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038.

Proisl, T. (2019). *The cooccurrence of linguistic structures*. Doctoral thesis, FAU.

Quan, L. (2011). Teaching Chinese for Specific Purposes and Its Textbook Compilation [J]. *Applied Linguistics*, 3.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Ramers, K. H. (2006). Topologische Felder: Nominalphrase und Satz im Deutschen. *Zeitschrift für Sprachwissenschaft*, 25(1):95–128.

Rausch, A. (2016). The contribution of complexity, accuracy and fluency to language for Specific Purposes. *Journal of Languages for Specific Purposes (JLSP)*, 29.

Redington, M., Crater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, 22(4):425–469.

Rehbein, I. (2010). Der Einfluss der Dependenzgrammatik auf die Computerlinguistik. *Zeitschrift für germanistische Linguistik*, 38(2):224–248.

Reis, M. (1980). On Justifying Topological Frames: 'Positional Field'and the Order of Nonverbal Constituents in German. *DRLAV. Documentation et Recherche en Linguistique Allemande Vincennes*, 22/23(1):59–85.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus. *Automatic treatment and analysis of learner corpus data*, 59:101–123.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., and Andreas, T. (2010). Das Falko-Handbuch: Korpusaufbau und Annotationen. *Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin*.

Richter, F. and Sailer, M. (2009). Phraseological clauses in constructional HPSG. In *Proceedings of the 16th international conference on Head-Driven Phrase Structure Grammar, university of Göttingen, germany*, pages 297–317.

Rivero, C. R. and Jamil, H. M. (2017). Efficient and scalable labeled subgraph matching using SGMatch. *Knowledge and Information Systems*, 51(1):61–87.

Robenalt, C. and Goldberg, A. E. (2016). Nonnative speakers do not take competing alternative expressions into account the way native speakers do. *Language Learning*, 66(1):60–93.

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*, volume 18. John Benjamins Publishing.

Römer, U., Roberson, A., O'Donnell, M. B., and Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, 38(1):115–135.

Rosch, E. (1983). Prototype classification and logical classification: The two systems. *New trends in conceptual representation: Challenges to Piaget's theory*, pages 73–86.

Roth, T. (2014). *Wortverbindungen und Verbindungen von Wörtern*, volume 94. BoD-Books on Demand.

Rousseau, F., Kiagias, E., and Vazirgiannis, M. (2015). Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1702–1712.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Ryabova, M. and Sergeychick, T. (2018). A Comparative Study of Approaches to Teaching French and English for Future Specialists in Mining Industry. In *E3S Web of Conferences*, volume 41, page 04039. EDP Sciences.

Sag, I. A. (2012). Sign-based construction grammar: An informal synopsis. *Sign-based construction grammar*, 193:69–202.

Sag, I. A., Boas, H. C., and Kay, P. (2012). Introducing sign-based construction grammar. *Sign-based construction grammar*, pages 1–30.

Saito, K. (to appear). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*.

Sánchez-López, L. (2018). Cultural and pragmatic aspects of L2 Spanish for academic purposes: new data on the current state of the question. *Journal of Spanish Language Teaching*, 5(2):102–114.

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., and Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition*, 13(2):231–252.

Saussure, F. d. (1916/1983). *Course in General Linguistics*. Duckworth, London. (trans. Roy Harris).

Schaub, M. T., Delvenne, J.-C., Rosvall, M., and Lambiotte, R. (2017). The many facets of community detection in complex networks. *Applied network science*, 2(1):4.

Scheinerman, E. R. and Ullman, D. H. (2011). *Fractional graph theory: a rational approach to the theory of graphs*. Courier Corporation.

Schieber, T. A., Carpi, L., Díaz-Guilera, A., Pardalos, P. M., Masoller, C., and Ravetti, M. G. (2017). Quantification of network structural dissimilarities. *Nature communications*, 8:13928.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, 66.

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? *Quantitative methods in cognitive semantics: Corpus-driven approaches*, pages 101–133.

Schmid, H.-J. (2015). A blueprint of the entrenchment-and-conventionalization model. *Yearbook of the German Cognitive Linguistics Association*, 3(1):3–26.

Schmid, H.-J. and Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*, volume 9. John Benjamins Publishing.

Schnakenberg, J. (1976). Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern physics*, 48(4):571.

Schneider, N. (2014). Lexical semantic analysis in natural language text. *Unpublished Doctoral Dissertation, Carnegie Mellon University*.

Schulz, P., Tracy, R., and Wenzel, R. (2008). Linguistische Sprachstandserhebung-Deutsch als Zweitsprache (LiSe-DaZ): Theoretische Grundlagen und erste Ergebnisse. In *Zweitspracherwerb. Diagnosen Verläufe Voraussetzungen; Beiträge aus dem 2. Workshop"Kinder mit Migrationshintergrund". Freiburg im Breisgau: Fillibach Verlag*, pages 17–41.

Seeker, W. and Çetinoğlu, Ö. (2015). A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Seretan, V. (2011). *Syntax-based collocation extraction*, volume 44. Springer Science & Business Media.

Seretan, V. and Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 401–410. IRIT Press.

Sharoff, S. (2017). Know thy corpus! Exploring frequency distributions in large corpora. *Essays in Honor of Adam Kilgarriff. Text Speech and Language Technology Series. Springer, Heidelberg*.

Sidtis, D., Canterucci, G., and Katsnelson, D. (2009). Effects of neurological damage on production of formulaic language. *Clinical linguistics & phonetics*, 23(4):270–284.

Simpson-Vlach, R. and Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4):487–512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus online only. 9 (1996), N. 1, 1996*, 9(1):1000–1032.

Sinclair, J. (1999). *The computer, the corpus and the theory of language*. EUT Edizioni Università di Trieste.

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

Siyanova-Chanturia, A. (2015). On the 'holistic'nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, 11(2):285–301.

Siyanova-Chanturia, A., Conklin, K., and Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

Snell-Hornby, M. (1983). *Verb-descriptivity in German and English: A contrastive study in semantic fields*. C. Winter Universitätsverlag.

Solé, R. V., Corominas-Murtra, B., Valverde, S., and Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26.

Starke, G. (1974). Das Bedeutungselement/+ distributiv/und sein Einfluß auf den Satzbau im Deutsehen—Ein Beitrag zur Untersuchung des Zusammenhangs zwischen Lexik und Grammatik. *STUF-Language Typology and Universals*, 27(1-3):231–238.

Steels, L. (2000). Language as a complex adaptive system. In *International Conference on Parallel Problem Solving from Nature*, pages 17–26. Springer.

Steels, L. (2013). Fluid construction grammar. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Steels, L. and Loetzsch, M. (2006). Perspective alignment in spatial language. *arXiv preprint cs/0605012*.

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19(3):513–531.

Stefanowitsch, A. (2011). Argument Structure: Item-Based or Distributed? *Zeitschrift für Anglistik und Amerikanistik*, 59(4):369–386.

Stefanowitsch, A. and Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Stefanowitsch, A. and Gries, S. T. (2005). Covarying collexemes. *Corpus linguistics and linguistic theory*, 1(1):1–43.

Stengers, H., Boers, F., Housen, A., and Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *IRAL-International Review of Applied Linguistics in Language Teaching*, 49(4):321–343.

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Steyer, K. (1998). Kollokationen als zentrales Übersetzungsproblem-Vorschläge für eine Kollokationsdatenbank Deutsch-Französisch/Französisch-Deutsch auf der Basis paralleler und vergleichbarer Korpora. In Bresson, D., editor, *Lexikologie und Lexikographie Deutsch-Französisch*, pages 95–113. Université Lumière.

Steyerberg, E. W. (2018). Validation in prediction research: the waste by data splitting. *Journal of clinical epidemiology*, 103:131–133.

Steyerberg, E. W. and Harrell, F. E. (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, 69:245–247.

Ströbel, M., Kerz, E., Wiechmann, D., and Neumann, S. (2016). CoCoGen-Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-Window Technique. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 23–31.

Ströbel, M., Kerz, E., Wiechmann, D., and Qiao, Y. (2018). Text Genre Classification Based on Linguistic Complexity Contours Using A Recurrent Neural Network. In *MRC@ IJCAI*, pages 56–63.

Swain, M. (1993). Second language testing and second language acquisition: is there a conflict with traditional psychometrics? *Language Testing*, 10(2):193–207.

Szudarski, P. (2017). Learning and Teaching L2 Collocations: Insights from Research. *TESL Canada Journal*, 34(3):205–216.

Szudarski, P. and Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics*, 26(2):245–265.

Tahmasebi, P. and Sahimi, M. (2016). Enhancing multiple-point geostatistical modeling: 1. Graph theory and pattern adjustment. *Water Resources Research*, 52(3):2074–2098.

Tanaka, J. W. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, 23(3):457–482.

Tao, H. and Chen, H. H.-J. (2019). *Chinese for Specific and Professional Purposes*. Springer.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental psychology*, 32(3):492.

Tardif, T., Shatz, M., and Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language*, 24(3):535–565.

Targońska, J. (2019). Kollokationskompetenz vs. Sprachfertigkeiten bzw. andere Sprachkompetenzen-ein Forschungsüberblick. *Glottodidactica. An International Journal of Applied Linguistics*, 46(1):179–196.

Tavakoli, P. and Uchihara, T. (to appear). To what extent are multiword sequences associated with oral fluency? *Language Learning*, 70(2).

Tesnière, L. (1965). *Eléments de syntaxe structurale*. éd. Klincksieck.

Thiele, S. (1990). Wandlungen in Wittgensteins Gebrauchstheorie der Bedeutung/Changes in Wittgenstein's theory of meaning as use. *Zeitschrift für germanistische Linguistik*, 18:127–149.

Titone, R. (1969). Guidelines for teaching a second language in its own environment. *The Modern Language Journal*, 53(5):306–309.

Tomasello, M. (1995). Language is not an instinct. *Cognitive Development*, 10:131–156.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4(4):156 – 163.

Tomasello, M. (2009). *Constructing a language*. Harvard university press.

Tomasello, M., Akhtar, N., Dodson, K., and Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language*, 24(2):373–387.

Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, 16(2):173–197.

Trinajstic, N. (2018). *Chemical graph theory*. Routledge.

Tucker, G. and Fawcett, R. (1996). So grammarians haven't the faintest idea?: Reconciling grammar and lexis in a systemic functional model of language. *Functional Descriptions: Theory in Practice*, pages 145–178.

Twain, M. (1880). *The awful German language*. BVK.

Ueffing, N., Och, F. J., and Ney, H. (2002). Generation of word graphs in statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 156–163.

Uhrig, P., Evert, S., and Proisl, T. (2018). Collocation candidate extraction from dependency-annotated corpora: exploring differences across parsers and dependency annotation schemes. In *Lexical Collocation Analysis*, pages 111–140. Springer.

Underwood, G., Schmitt, N., and Galpin, A. (2004). The eyes have it. *Formulaic sequences: Acquisition, processing, and use*, 9:153.

Valsecchi, M., Künstler, V., Saage, S., White, B. J., Mukherjee, J., and Gegenfurtner, K. R. (2014). Advantage in reading lexical bundles is reduced in non-native speakers. *Journal of Eye Movement Research*, 6(5):1–15.

Van Lancker, D. (1988). Nonpropositional speech: Neurolinguistic studies. In *Progress in the psychology of language*, pages 49–118. Lawrence Erlbaum Associates.

Vassiljev, L., Skopinskaja, L., and Liiv, S. (2015). The treatment of lexical collocations in EFL coursebooks in the Estonian secondary school context. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 11:297–311.

Veltkamp, G. M., Recio, G., Jacobs, A. M., and Conrad, M. (2013). Is personality modulated by language? *International Journal of Bilingualism*, 17(4):496–504.

Venkatapathy, S. and Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906. Association for Computational Linguistics.

Verbrugge, R. J. (2006). Nonergodic corruption dynamics (or, why do some regions within a country become more corrupt than others?). *Journal of Public Economic Theory*, 8(2):219–245.

Veronis, J. and Ide, N. M. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 389–394. Association for Computational Linguistics.

Vyatkina, N. (2015). New developments in the study of L2 writing complexity: An editorial. *Journal of Second Language Writing*, 29:1–2.

Vyatkina, N. (2016). The Kansas Developmental Learner corpus (KANDEL). *International Journal of Learner Corpus Research*, 2(1):101–119.

Vyatkina, N., Hirschmann, H., and Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, 29:28–50.

Wachs-Lopes, G. A. and Rodrigues, P. S. (2016). Analyzing natural human language from the point of view of dynamic of a complex network. *Expert Systems with Applications*, 45:8–22.

Wallace, R. (2013). A new formal approach to evolutionary processes in socioeconomic systems. *Journal of Evolutionary Economics*, 23(1):1–15.

Wallner, F. (2014). Kollokationen in Wissenschaftssprachen. *Zur lernerlexikographischen Relevanz ihrer wissenschaftssprachlichen Gebrauchsspezifika. Tübingen: Stauffenburg*, pages 382–384.

Wan, S. (in prep.). *Argumentationsstrategien von chinesischen Deutschlernern*. Univ. diss., Humboldt-Universität zu Berlin.

Warren, J. (2017). Revisiting Quine on truth by convention. *Journal of Philosophical Logic*, 46(2):119–139.

Webber, J. and Robinson, I. (2018). *A programmatic introduction to neo4j*. Addison-Wesley Professional.

Weigel, A. V., Simon, B., Tamkun, M. M., and Krapf, D. (2011). Ergodic and nonergodic processes coexist in the plasma membrane as observed by single-molecule tracking. *Proceedings of the National Academy of Sciences*, 108(16):6438–6443.

Welke, K. (2011). *Valenzgrammatik des Deutschen: Eine Einführung*. Walter de Gruyter.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2):253–290.

Wiener, S., Speer, S. R., and Shank, C. (2012). Effects of frequency, repetition and prosodic location on ambiguous Mandarin word production. In *Speech Prosody 2012*, pages 528–531.

Wilcke, W., de Boer, V., de Kleijn, M., van Harmelen, F., and Scholten, H. (2018). User-centric pattern mining on knowledge graphs: An archaeological case study. *Journal of Web Semantics*, page 100486.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., and Dodds, P. S. (2015). Zipf's law holds for phrases, not words. *Nature Scientific Reports*, 5:12209.

Wisniewski, K. (2017a). Empirical learner language and the levels of the common European framework of reference. *Language Learning*, 67(S1):232–253.

Wisniewski, K. (2017b). The Empirical Validity of the Common European Framework of Reference Scales. An Exemplary Study for the Vocabulary and Fluency Scales in a Language Testing Context. *Applied Linguistics*, 39(6):933–959.

Wittenburg, K. B. (1986). *Natural Language Parsing with Combinatory Categorical Grammar in a Graph-unification-based Formalism*. PhD thesis, University of Texas at Austin.

Wittgenstein, L. (1953). *Ludwig Wittgenstein Werkausgabe*, volume "Band 1". Suhrkamp, Frankfurt am Main.

Wonnacott, E., Brown, H., and Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, 95:36 – 48.

Wonnacott, E., Newport, E. L., and Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive psychology*, 56(3):165–209.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. Bloomsbury Publishing.

Wood, D. C. and Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes*, 15:1–13.

Wood, S., N., Pya, and S"afken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111:1548–1575.

Woods, W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10):591–606.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32:231–254.

Wray, A. (2013). Formulaic language. *Language Teaching*, 46(3):316–334.

Wray, A. and Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, 20(1):1–28.

Wu, S., Witten, I. H., and Franken, M. (2010). Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge. *ReCALL*, 22(1):83–102.

Wulff, S. (2006). Go-V vs. go-and-V in English: A case of constructional synonymy. *Corpora in Cognitive Linguistics. Corpus-based approaches to syntax and lexis*, pages 101–126.

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach*. A&C Black.

Xu, P. and Jelinek, F. (2004). Random Forests in Language Modeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 325–332.

Xue, L. (2015). Delexicalizing features analysis of mandarin Chinese light verbs. In *2015 International Conference on Social Science, Education Management and Sports Education*. Atlantis Press.

Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. *Bilingualism across the lifespan*, pages 55–72.

Young, R. (1995). Discontinuous interlanguage development and its implications for oral proficiency rating scales. *Applied Language Learning*, 6:13–26.

Zaprudski, S. (2007). In the grip of replacive bilingualism: the Belarusian language in contact with Russian. *International Journal of the Sociology of Language*, 2007(183):97–118.

Zeldes, A. (2012). *Productivity in argument selection: From morphology to syntax*, volume 260. Walter de Gruyter.

Zeldes, A. (2013a). Komposition als Konstruktionsnetzwerk im fortgeschrittenen L2-Deutsch. *Zeitschrift für germanistische Linguistik*, 41(2):240–276.

Zeldes, A. (2013b). Productive argument selection: Is lexical semantics enough? *Corpus Linguistics and Linguistic Theory*, 9(2):263–291.

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., and Skiba, D. (2012). Das Wissenschaftliche Netzwerk" Kobalt-DaF". *Zeitschrift für germanistische Linguistik*, 40(3):457–458.

Zipf, G. K. (1965). The Psycho-Biology of Language. An Introduction to Dynamic Philology. 1935.